A Test for Superior Predictive Ability

Stanford University, Department of Economics, 579 Serra Mall, Stanford, CA 94305 (peterhan@stanford.edu)

We propose a new test for superior predictive ability. The new test compares favorably to the reality check (RC) for data snooping, because it is more powerful and less sensitive to poor and irrelevant alternatives. The improvements are achieved by two modifications of the RC. We use a studentized test statistic that reduces the influence of erratic forecasts and invoke a sample-dependent null distribution. The advantages of the new test are confirmed by Monte Carlo experiments and an empirical exercise in which we compare a large number of regression-based forecasts of annual U.S. inflation to a simple random-walk forecast. The random-walk forecast is found to be inferior to regression-based forecasts and, interestingly, the best sample performance is achieved by models that have a Phillips curve structure.

KEY WORDS: Forecast evaluation; Forecasting; Inequality testing; Multiple comparison; Testing for superior predictive ability.

1. INTRODUCTION

Testing whether a particular forecasting procedure is outperformed by alternative forecasts represents a test of *superior predictive ability* (SPA). White (2000) developed a framework for comparing multiple forecasting models and proposed a test for SPA that is known as the reality check (RC) for data snooping. Here the term "model" is used in a broad sense that includes forecasting rules/methods, which need not involve modeling data. In White's framework, m alternative forecasts (where m is a fixed number) are compared with a benchmark forecast, where the predictive abilities are defined by expected loss. The complexity of this inference problem arises from the need to control for the full set of alternatives.

In this article, we propose a new test for SPA. Our framework is identical to that of White (2000), but we take a different path in our construction of the test. To be specific, we use a different test statistic and invoke a *sample-dependent distribution* under the null hypothesis. Compared with the RC, the new test is more powerful and less sensitive to the inclusion of poor and irrelevant alternatives.

We make three contributions in this article. First, we provide a theoretical analysis of the testing problem and expose some of its important aspects. Our theoretical results reveal that the RC can be manipulated by the including poor and irrelevant forecasts in the set of alternative forecasts. This problem is alleviated by studentizing the test statistic and by invoking a sample-dependent null distribution. The latter is based on a novel procedure that incorporates additional sample information to identify the "relevant" alternatives. Second, we provide a detailed explanation of a bootstrap implementation of our test for SPA. Third, we apply the tests in an empirical analysis of U.S. inflation. Our benchmark is a simple random-walk forecast that uses current inflation as the prediction of future inflation. The benchmark is compared with a large number of regression-based forecasts, and our empirical results show that the benchmark is significantly outperformed. Interestingly, the strongest evidence is provided by regression models that have a Phillips curve structure.

When testing for SPA, the question of interest is whether any alternative forecast is better than the benchmark forecast or, equivalently, whether the best alternative forecasting model is better than the benchmark. This question can be addressed by testing the null hypothesis that "the benchmark is not inferior to any alternative forecast." This testing problem is relevant for applied econometrics, because several ideas and specifications are often used before a model is selected. This *mining* over alternative forecasts may be exacerbated if more than one researcher is searching for a good forecasting model. (For a more complete discussion of this issue, see Sullivan, Timmermann, and White 2003 and references therein.) Testing for SPA is useful for a forecaster who wants to explore whether a better forecasting model than the model currently being used to make predictions is available. After a search over several alternative forecasts, the relevant question is whether one of the alternative forecasts is significantly more accurate than the benchmark. (The test for SPA can also be used to test an economic theory that places restrictions on the predictability of certain variables, such as the *efficient markets hypothesis*; see Sullivian et al. 1999.)

Tests for equal predictive ability (EPA) in a general setting were proposed by Diebold and Mariano (1995) and West (1996), where the framework of West can accommodate the situation where forecasts involve estimated parameters. Harvey, Leybourne, and Newbold (1997) suggested a modification of the Diebold–Mariano test that leads to better small-sample properties. A test for comparing multiple-nested models was given by Harvey and Newbold (2000), and McCracken (2000) derived results for the case with estimated parameters and nondifferentiable loss functions, such as the mean absolute deviation loss function. West and McCracken (1998) developed regression-based tests, and other extensions were made by Harvey et al. (1998), Chao, Corradi, and Swanson (2001), Clark and McCracken (2001), West (2001), and Corradi and Swanson (2002), who considered tests for forecast encompassing, and by Corradi, Swanson, and Olivetti (2001), who compared forecasting models that include cointegrated variables. (For a discussion of in-sample versus out-of-sample testing, see Inoue and Kilian 2004.)

Whereas the frameworks of Diebold and Mariano (1995) and West (1996) involve tests for EPA, the testing problem in White's framework is a test for SPA. The distinction is important because the former leads to a simple null hypothesis, whereas the latter leads to a composite hypothesis. One of the

Journal of Business & Economic Statistics
October 2005, Vol. 23, No. 4
DOI 10.1198/073500105000000063

main complications in composite hypotheses testing is that (asymptotic) distributions typically depend on nuisance parameters, such that the null distribution is not unique. The usual way to handle this ambiguity is to use the *least favorable configuration* (LFC), which is sometimes referred to as "the point least favorable to the alternative." Our analysis shows that the LFC-based approach leads to some rather unfortunate properties when testing for SPA. The following situation delivers key insight to the advantages of using a sample-dependent null distribution. Let $p_{\min}$ denote the smallest p value of the m pairwise comparisons (comparing each alternative with the benchmark); then the Bonferroni bound test (at level α) rejects the null hypothesis if $p_{\min} < \alpha/m$. It is now evident that the power of this test can be driven to 0 by adding poor and irrelevant alternatives to the comparison, because this increases m but does not affect $p_{\min}$. However, sample information will (at least asymptotically) identify the poor and irrelevant alternative, which allows us to use a smaller denominator when defining the critical value, for example, α/m_0 for some $m_0 \leq m$. Although our testing procedure is quite different from the conservative Bonferroni bound test, our sample-dependent null distribution is similar to this improvement of the Bonferroni bound test, although the (presumed) poor alternatives are not discarded entirely in our framework.

In relation to the existing literature on forecast evaluation and comparison, it is important to acknowledge a limitation of the specific test that we propose in this article. A comparison of models with parameters that are estimated recursively is not accommodated by our framework, because this situation violates our stationarity assumption. (For recent progress on this problem in the present context, see Corradi and Swanson 2005a.) However, our framework does permit parameters that are estimated once (fixed scheme) or with a moving window (rolling schemes), as we discuss in Section 2. The advantages of the studentized test statistic and our sample-dependent null distribution do not rely on stationarity, so these modifications are expected to be useful in a more general context. A related issue concerns the optimality of our test. Although the new test dominates the RC, we do not claim that it is optimal. The lack of an optimality result is not surprising, because such results are rare in composite hypothesis testing. It is also worth observing that leading statisticians continue to quarrel about what constitutes a suitable test in this context (see Perlman and Wu 1999 and the comments on that article by Berger, Cox, McDermott, and Wang).

This article is organized as follows. Section 2 introduces the new test for SPA and contains our theoretical results. Section 3 provides the details of the bootstrap implementation. Section 4 contains a simulation-based study of the finite-sample properties of the new test for SPA and compares it with those of the RC. Section 5 contains an empirical forecasting exercise of U.S. inflation, and Section 6 gives a summery and some concluding remarks. All proofs are presented in an Appendix.

2. TESTING FOR SUPERIOR PREDICTIVE ABILITY

We consider a situation where a decision must be made h periods in advance and let $\{\delta_{k,t-h}, k = 0, 1, \ldots, m\}$ be a finite set of possible decision rules. Decisions are evaluated with a real-valued loss function, $L(\xi_t, \delta_{k,t-h})$, where ξ_t is a random variable that represents the aspects of the decision problem that are unknown at the time that the decision is made. We evaluate forecasts in terms of their expected loss, $E[L(\xi_t, \delta_{k,t-h})]$. Thus we need not assume that any of the forecasts are constructed from a correctly specified model. Whenever $\delta_{k,t-h} = \delta_{k,t-h}(\hat{\theta}_{k,t-h})$ is based on estimated parameters, $\hat{\theta}_{k,t-h}$, these are likely to influence the expected loss—typically by increasing the expected loss. We make assumptions that do not permit parameters that are estimated with the recursive scheme. However, the rolling scheme is accommodated by our framework, and so is the fixed scheme when the comparison of forecasts is interpreted as being conditional on the estimated parameters. An overview of our notation is given in Table 1. This provides a general framework for comparing forecasts and decision rules. Our leading example is the comparison of forecasts, so we often refer to $\delta_{k,t-h}$ as the kth forecasting model. The first model, $k = 0$, has a special role and is referred to as the benchmark. The decision rule, $\delta_{k,t-h}$, can represent a point forecast, an interval forecast, a density forecasts, or a trading rule for an investor, as we illustrate next with some examples.

Example 1 (Point forecast). Let $\delta_{k,t-h}$, $k = 0, 1, \ldots, m$, be different point forecasts of a real random variable ξ_t. The mean squared error loss function, $L(\xi_t, \delta_{k,t-h}) = (\xi_t - \delta_{k,t-h})^2$, is an example of a loss function that could be used to compare the different forecasts.

Example 2 (Conditional distribution and value-at-risk forecasts). Let ξ_t be a conditional density on $\mathbb{R}$, and let $\delta_{k,t-h}$ be a forecast of ξ_t. Then we might evaluate the precision of δ_k by the Kolmogorov–Smirnov statistic, $L(\xi_t, \delta_{k,t-h}) = \sup_{x\in\mathbb{R}} |\int_{-\infty}^{x} [\xi_t(y) - \delta_{k,t-h}(y)]\, dy|$, or a Kullback–Leibler measure, $L(\xi_t, \delta_{k,t-h}) = \int_{-\infty}^{\infty} \log[\delta_{k,t-h}(x)/\xi_t(x)]\xi_t(x)\, dx$. Alternatively, $\delta_{k,t-h}$ could be a value-at-risk measure (at quantile α) that may be evaluated with $L(\xi_t, \delta_{k,t-h}) = |\int_{-\infty}^{\delta_{k,t-h}} \xi_t(x)\, dx - \alpha|$.

In Example 2, ξ_t will often be unobserved, which creates additional complications for empirical evaluation and comparison. When a proxy is substituted for ξ_t it can cause the empirical ranking of alternatives to be inconsistent for the intended (true) ranking (see Hansen and Lunde 2005a). Corradi and Swanson (2005b) recently derived an RC-type test for comparing conditional density forecasts, which is closely related to the problem of Example 2. Their test is similar to that of White (2000), because their test statistic is also the maximum of multiple nonstudentized quantities. So it would be interesting to analyze whether our two modifications can be implemented in their framework.

Example 3 (Trading rules). Let $\delta_{k,t-1}$ be a binary variable that instructs a trader to take either a short ($\delta = -1$) or a long ($\delta = 1$) position in an asset at time $t - 1$. The kth *trading rule* yields the profit $\pi_{k,t} = \delta_{k,t-1} r_t$, where r_t is the return on the asset in period t. A trader who is currently using the rule, δ_0, might be interested to know whether an alternative rule has a larger expected profit than δ_0. This can be formulated in our framework by setting $\xi_t = r_t$ and $L(\xi_t, \delta_{k,t-1}) = -\delta_{k,t-1}\xi_t$.

Table 1. Overview of Notation and Definitions

$t = 1, \ldots, n$	Sample period for the model comparison
$k = 0, 1, \ldots, m$	Model index ($k = 0$ is the benchmark)
ξ_t	Object (variable) of interest
$\delta_{k,t-h}$	The kth decision rule (e.g., h-step-ahead forecast of ξ_t)
$L_{k,t} \equiv L(\xi_t, \delta_{k,t-h})$	Observed loss of the kth decision rule/forecast
$d_{k,t} \equiv L_{0,t} - L_{k,t}$	Performance of model k relative to the benchmark
$\bar{d}_k \equiv n^{-1} \sum_{t=1}^{n} d_{k,t}$	Average relative performance of model k
$\mathbf{d}_t \equiv (d_{1,t}, \ldots, d_{m,t})'$	Vector of relative performances at time t
$\bar{\mathbf{d}} \equiv n^{-1} \sum_{t=1}^{n} \mathbf{d}_t$	Vector of average relative performance
$\mu_k \equiv E(d_{k,t})$	Expected excess performance of model k
$\boldsymbol{\mu} \equiv (\mu_1, \ldots, \mu_m)'$	Vector of expected excess performances
$\boldsymbol{\Omega} \equiv \text{avar}(n^{1/2}\bar{\mathbf{d}})$	Asymptotic $m \times m$ covariance matrix

The benchmark in Example 3 could be $\delta_{0,t} = 1$, which is the rule that is always "long in the market." This was the benchmark used by Sullivan et al. (1999, 2001) who evaluated the significance of technical trading rules and calendar effects in stock returns.

2.1 Hypothesis of Interest

We are interested to know whether any of the models, $k = 1, \ldots, m$, are better than the benchmark in terms of expected loss. So we seek a test of the null hypothesis that *the benchmark is not inferior to any of the alternatives*. The variables that are key for our analysis are the relative performance variables, which are defined by

$$d_{k,t} \equiv L(\xi_t, \delta_{0,t-h}) - L(\xi_t, \delta_{k,t-h}), \qquad k = 1, \ldots, m.$$

So $d_{k,t}$ denotes the performance of model k relative to the benchmark at time t, and we stack these variables into the vector of relative performances, $\mathbf{d}_t = (d_{1,t}, \ldots, d_{m,t})'$. Provided that $\boldsymbol{\mu} \equiv E(\mathbf{d}_t)$ is well defined, we can now formulate the null hypothesis of interest as

$$H_0 : \boldsymbol{\mu} \leq \mathbf{0}, \tag{1}$$

and our maintained hypothesis is $\boldsymbol{\mu} \in \mathbb{R}^m$.

We work under the assumption that model k is better than the benchmark if and only if $E(d_{k,t}) > 0$. So we focus exclusively on the properties of $\mathbf{d}_t$ and abstract entirely from all aspects that relate to the construction of the δ-variables. Thus $\mathbf{d}_t$, $t = 1, \ldots, n$, is de facto viewed as our data, and we therefore state all assumptions in terms $\mathbf{d}_t$. Specifically we make the following assumption.

Assumption 1. The vector of relative loss variables, $\{\mathbf{d}_t\}$, is (strictly) stationary and α-mixing of size $-(2+\delta)(r+\delta)/(r-2)$, for some $r > 2$ and $\delta > 0$, where $E|\mathbf{d}_t|^{r+\delta} < \infty$ and $\text{var}(d_{k,t}) > 0$ for all $k = 1, \ldots, m$.

Assumption 1 is made for two reasons: first, to ensure that certain population moments such as $\boldsymbol{\mu}$ are well defined, and second, to justify the use of bootstrap techniques that we describe in detail in Section 3. Note that Assumption 1 does not require that the individual loss variables, $L(\xi_t, \delta_{k,t-h})$, be stationary. An immediate consequence of Assumption 1 is that a central limit theorem applies, such that

$$n^{1/2}(\bar{\mathbf{d}} - \boldsymbol{\mu}) \xrightarrow{d} \text{N}_m(\mathbf{0}, \boldsymbol{\Omega}), \tag{2}$$

where $\bar{\mathbf{d}} \equiv n^{-1} \sum_{t=1}^{n} \mathbf{d}_t$ and $\boldsymbol{\Omega} \equiv \text{avar}(n^{1/2}(\bar{\mathbf{d}} - \boldsymbol{\mu}))$ (see, e.g., de Jong 1997).

Diebold and Mariano (1995) and West (1996) provided sufficient conditions that also lead to the asymptotic normality in (2). Giacomini and White (2003) established this property for a related testing problem. However, the asymptotic normality does not hold in general. An important exception is the situation where the benchmark is nested in all alternative models (under the null hypothesis) and the parameters are estimated recursively. In this situation the limiting distribution will typically be given as a function of Brownian motions (see, e.g., Clark and McCracken 2001). When comparing nested models, the null hypothesis simplifies to the simple hypothesis, $\boldsymbol{\mu} = \mathbf{0}$. So in this case it seems more appropriate to apply a test for EPA, such as that of Harvey and Newbold (2000), which can be used to compare multiple-nested models.

At this point, all essential aspects of our framework are identical to those of White (2000). White proceeded by constructing the RC from the test statistic,

$$T_n^{\text{RC}} \equiv \max\left(n^{1/2}\bar{d}_1, \ldots, n^{1/2}\bar{d}_m\right),$$

and an asymptotic null distribution based on $n^{1/2}\bar{\mathbf{d}} \sim \text{N}_m(\mathbf{0}, \hat{\boldsymbol{\Omega}})$, where $\hat{\boldsymbol{\Omega}}$ is a consistent estimator of $\boldsymbol{\Omega}$. Here it is worth noting that the RC relies on an asymptotic null distribution that assumes $\mu_k = 0$ for all k, even though all negative values of μ_k also conform with the null hypothesis. This aspect is the underlying topic of Sections 2.3 and 2.4, but first we discuss a studentization of the test statistic.

Given the asymptotic normality of $\bar{\mathbf{d}}$, it may seem natural to use a quadratic-form test statistic to test H_0, such as the likelihood ratio test used by Wolak (1987). However, the situation that we have in mind is one in which m is too large to obtain a sensible estimate of all elements of $\boldsymbol{\Omega}$. Instead we consider simpler statistics, such as T_n^{SPA} (defined later) that requires only that the diagonal elements of $\boldsymbol{\Omega}$ be estimated. It is not surprising that nonquadratic statistics will be nonpivotal—even asymptotically—because their asymptotic distribution will depend on (some elements of) the covariance matrix, which makes $\boldsymbol{\Omega}$ a nuisance parameter. To handle this problem, we follow White (2000) and use a bootstrap method that implicitly takes care of this nuisance parameter problem. So our motivation for using the bootstrap is not driven by higher-order refine-

ments, but is merely to handle this nuisance parameter problem.

We analyze this testing problem in the remainder of this section, and our findings motivate the following two recommendations that spell out the differences between the RC and our new test for SPA:

1. Use the studentized test statistic,

$$T_n^{\mathrm{SPA}} \equiv \max\left[\max_{k=1,\ldots,m} \frac{n^{1/2}\bar{d}_k}{\hat{\omega}_k}, 0\right],$$

where $\hat{\omega}_k^2$ is some consistent estimator of $\omega_k^2 \equiv \mathrm{var}(n^{1/2}\bar{d}_k)$.

2. Invoke a null distribution that is based on $\mathrm{N}_m(\hat{\boldsymbol{\mu}}^c, \hat{\boldsymbol{\Omega}})$, where $\hat{\boldsymbol{\mu}}^c$ is a carefully chosen estimator for $\boldsymbol{\mu}$ that conforms with the null hypothesis. Specifically, we suggest the estimator

$$\hat{\mu}_k^c = \bar{d}_k \mathbb{1}_{\{n^{1/2}\bar{d}_k/\hat{\omega}_k \le -\sqrt{2\log\log n}\}}, \qquad k = 1, \ldots, m,$$

where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function.

We explain our reasons for this choice of $\boldsymbol{\mu}$-estimator in Section 2.4, but it is important to understand that using a consistent estimator of $\boldsymbol{\mu}$ need not produce a valid test.

2.2 Choice of Test Statistic

When the benchmark has the best sample performance ($\bar{\mathbf{d}} \le \mathbf{0}$), the test statistic is normalized to 0. In this case there is no evidence against the null hypothesis, and consequently the null should not be rejected. The normalization is convenient for theoretical reasons, because we avoid a divergence problem (to $-\infty$) that would otherwise occur when $\boldsymbol{\mu} < \mathbf{0}$.

As discussed in Section 1, there are few optimality results in the context of composite hypothesis testing. This is particularly the case for the present problem of testing multiple inequalities. However, some arguments that justify our choice of test statistic T_n^{SPA} (instead of T_n^{RC}) are called on. Although we argue that T_n^{SPA} is preferable to T_n^{RC}, it cannot be shown that the former uniformly dominates the latter in terms of power. In fact, there are situations where T_n^{RC} leads to a more powerful test (such as the case where $\omega_j^2 = \omega_k^2\ \forall j, k = 1, \ldots, m$). However, such exceptions are unlikely to be of much empirical relevance, as we discuss later. So we are comfortable recommending the use of T_n^{SPA} in practice, and it is worth pointing out that studentization of the individual statistics is the conventional approach to multiple comparisons (see Miller 1981; Savin 1984). This studentization is also embedded in the related approach where the individual statistics are converted into "p values," with the smallest p value used as the test statistic (see Tippett 1931; Folks 1984; Marden 1985; Westfall and Young 1993; Dufour and Khalaf 2002). In the present context, Romano and Wolf (2005) also adopted the studentized test statistic (see also Lehmann and Romano 2005, chap. 9).

Our main argument for studentization is that it typically will improve the power. This can be understood from the following simple example.

Example 4. Consider the case where $m = 2$ and suppose that

$$n^{1/2}(\bar{\mathbf{d}} - \boldsymbol{\mu}) \sim \mathrm{N}_2\left(\mathbf{0}, \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}\right),$$

where the covariance is 0 (a simplification that is not necessary for our argument). Now consider the particular local alternative where $\mu_2 = 2n^{-1/2} > 0$. Here $\bar{d}_2$ is expected to yield a fair amount of evidence against $H_0 : \boldsymbol{\mu} \le \mathbf{0}$, because the t-statistic, $n^{1/2}\bar{d}_2/\hat{\omega}_k$, will be centered about 2. It follows that the null distributions (using $\boldsymbol{\mu} = \mathbf{0}$) are given by $T_n^{\mathrm{RC}} \sim F_0(x) \equiv \Phi(x/2)\Phi(x)$ and $T_n^{\mathrm{SPA}} \overset{a}{\sim} G_0(x) \equiv \Phi(x)\Phi(x)$, whereas $T_n^{\mathrm{RC}} \sim F_1(x) \equiv \Phi(x/2)\Phi(x+2)$ and $T_n^{\mathrm{SPA}} \overset{a}{\sim} G_1(x) \equiv \Phi(x)\Phi(x+2)$ under the local alternative. Here $\Phi(\cdot)$ denotes the standard Gaussian distribution and $\overset{a}{\sim}$ means "asymptotically distributed as." Figure 1 shows the upper tails of the null distributions, $1 - F_0(x)$ and $1 - G_0(x)$ (thick lines) and the upper tails of $1 - F_1(x)$ and $1 - G_1(x)$ (thin lines) that represent the distributions of the test statistics under the local alternative. Dotted

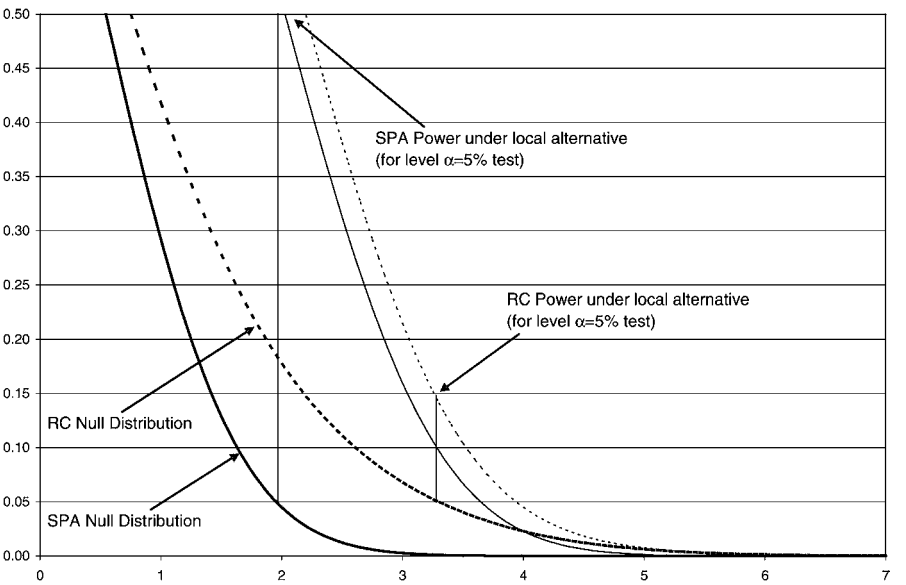

Figure 1. (One minus) The cdfs for the Test Statistics T^{RC} and T^{SPA} Under the Null Hypothesis, $\mu_1 = \mu_2 = 0$, and the Local Alternative, $\mu_2 = 2/\sqrt{n} > 0$. [- - - 1 − F0(x); ····· 1 − F1(x); —— 1 − G0(x); —— 1 − G1(x).] Studentization improves the power from about 15% to about 53%.

lines represent for the distributions of T_n^{RC}, and solid lines represent for the distributions of T_n^{SPA}. The power for a given level of either of the two tests can be read off the figure, and we have singled out the powers of the 5%-level tests. These reveal that studentization more than triples the power, from about 15% to about 53%. So the RC is much less likely to detect the false null, because the noisy $\bar{d}_1$ conceals the evidence against H_0 that $\bar{d}_2$ provides.

The preceding example highlights the advantages of studentizing the individual statistics, because it avoids a comparison of objects measured in different "units of standard deviation" (avoiding a comparison of apples and bananas). There is one exception where studentization may reduce the power, when the best performing model is associated with the largest variance [i.e., if $\mathrm{var}(\bar{d}_2) \geq \mathrm{var}(\bar{d}_1)$ in the previous example]. We consider this case to be of little empirical relevance, because poorly performing models also tend to have the most erratic performances in practice. Moreover, the loss in power from estimating ω_k^2, $k = 1, \ldots, m$, is quite modest when these are estimated precisely, as is the case when n is large.

In the remainder of this section we formulate our theoretical results that motivate our data-dependent choice of null distribution. We derive our results for a broad class of test statistics to emphasize that our results are not specific to the two statistics, T_n^{RC} and T_n^{SPA}. This is also convenient because other statistics (from this class of statistics) may be used in future applied work.

2.3 Theoretical Results for a Class of Test Statistics

We consider a class of test statistics, where each of the statistics satisfies the following conditions.

Assumption 2. The test statistic has the form $T_n = \varphi(\mathbf{U}_n, \mathbf{V}_n)$, where $\mathbf{U}_n \equiv n^{1/2}\bar{\mathbf{d}}$ and $\mathbf{V}_n \xrightarrow{p} \mathbf{v}_0 \in \mathbb{R}^q$ (a constant). The mapping, $\varphi(\mathbf{u}, \mathbf{v})$, is continuous in $\mathbf{u}$ on $\mathbb{R}^m$ and continuous in $\mathbf{v}$ in a neighborhood of $\mathbf{v}_0$. Further, φ has the following properties:

(a) $\varphi(\mathbf{u}, \mathbf{v}) \geq 0$ and $\varphi(\mathbf{0}, \mathbf{v}) = 0$.

(b) $\varphi(\mathbf{u}, \mathbf{v}) = \varphi(\mathbf{u}^+, \mathbf{v})$, where $u_k^+ = \max(0, u_k)$, $k = 1, \ldots, m$.

(c) $\varphi(\mathbf{u}, \mathbf{v}) \to \infty$, if $u_k \to \infty$ for some $k = 1, \ldots, m$.

Thus, in addition to the sample average, $\bar{\mathbf{d}}$, the test statistic may depend on the data through $\mathbf{V}_n \equiv v(\mathbf{d}_1, \ldots, \mathbf{d}_n)$, as long as $\mathbf{V}_n$ converges in probability to a constant (or vector of constants). Assumption 2(a) is a normalization (if $\bar{\mathbf{d}} = \mathbf{0}$, then there is no evidence against H_0), Assumption 2(b) states that only the positive elements of $\mathbf{u}$ matter for the value of the test statistic, and Assumption 2(c) requires that the test statistic diverges to infinity as the evidence against the null hypothesis increases (to infinity).

The mapping $(\boldsymbol{\mu}, \boldsymbol{\Omega}) \mapsto \boldsymbol{\Omega}^0$, given by

$$\Omega_{ij}^0 \equiv \Omega_{ij} \mathbb{1}_{\{\mu_i = \mu_j = 0\}}, \qquad i, j = 1, \ldots, m,$$

defines an $m \times m$ covariance matrix, $\boldsymbol{\Omega}^0$, that plays a role in our asymptotic results. So $\boldsymbol{\Omega}^0$ is similar to $\boldsymbol{\Omega}$, except that the elements of certain rows and columns have been set to 0. An example of how $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ translate into $\boldsymbol{\Omega}^0$ is as follows:

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ -2 \\ 0 \end{pmatrix}, \qquad \boldsymbol{\Omega} = \begin{pmatrix} \omega_{11}^2 & \omega_{12} & \omega_{13} \\ \omega_{21} & \omega_{22}^2 & \omega_{23} \\ \omega_{31} & \omega_{32} & \omega_{33}^2 \end{pmatrix},$$

$$\boldsymbol{\Omega}^0 = \begin{pmatrix} \omega_{11}^2 & 0 & \omega_{13} \\ 0 & 0 & 0 \\ \omega_{31} & 0 & \omega_{33}^2 \end{pmatrix},$$

and $\boldsymbol{\Omega}^0$ has at most rank m_0, where m_0 is the number of elements in $\boldsymbol{\mu}$ that equal 0.

The following theorem provides the asymptotic null distribution for all test statistics that satisfy Assumption 2.

Theorem 1. Suppose that Assumptions 1 and 2 hold and let F_0 be the cumulative distribution function (cdf) of $\varphi(\mathbf{Z}, \mathbf{v}_0)$, where $\mathbf{Z} \sim \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Omega}^0)$. Under the null hypothesis, $\boldsymbol{\mu} \leq \mathbf{0}$, we have that $\varphi(n^{1/2}\bar{\mathbf{d}}, \mathbf{V}_n) \xrightarrow{d} F_0$, where $\mathbf{v}_0 = \mathrm{plim}\mathbf{V}_n$. Under the alternative, $\boldsymbol{\mu} \not\leq \mathbf{0}$, we have that $\varphi(n^{1/2}\bar{\mathbf{d}}, \mathbf{V}_n) \xrightarrow{p} \infty$.

The test statistic T_n^{SPA} satisfies Assumption 2, whereas that of the RC does not. It is nevertheless possible to obtain critical values for T_n^{RC} from Theorem 1. This is done by applying Theorem 1 to the test statistic $T_n^{\mathrm{RC}^+} = \max(T_n^{\mathrm{RC}}, 0)$ that satisfies Assumption 2 and noting that the distributions of $T_n^{\mathrm{RC}^+}$ and T_n^{RC} coincide on the positive axis, which is the relevant support for the critical value. Alternatively, the asymptotic distribution of T_n^{RC} can be obtained directly, as we do in the following corollary.

Corollary 1. Let $m_0 \leq m$ be the number of models with $\mu_k = 0$, define $\boldsymbol{\Sigma}$ to be the $m_0 \times m_0$ submatrix of $\boldsymbol{\Omega}$ that contains the (i, j)th element of $\boldsymbol{\Omega}$ if $\mu_i = \mu_j = 0$, and let $\zeta_{\boldsymbol{\Sigma}}$ denote the distribution of $Z_{\max} \equiv \max_{j=1,\ldots,m_0} Z_j^0$, where $\mathbf{Z}^0 = (Z_1^0, \ldots, Z_{m_0}^0)' \sim \mathrm{N}_{m_0}(\mathbf{0}, \boldsymbol{\Sigma})$. Then $T_n^{\mathrm{RC}} \xrightarrow{d} \zeta_{\boldsymbol{\Sigma}}$ if $\max_k \mu_k = 0$, whereas $T_n^{\mathrm{RC}} \xrightarrow{p} -\infty$ if $\mu_k < 0$ for all $k = 1, \ldots, m$. Under the alternative where $\mu_k > 0$ for some k, it holds that $T_n^{\mathrm{RC}} \xrightarrow{p} \infty$.

Theorem 1 and Corollary 1 demonstrate that it is only the binding constraints (i.e., those with $\mu_k = 0$) that matter for the asymptotic distribution. Naturally, the number of binding constraints can be small relative to the number of inequalities, m, being tested. This result is known from the problem of testing linear inequalities in linear (regression) models (see Perlman 1969; Wolak 1987, 1989b; Robertson, Wright, and Dykstra 1988; Dufour 1989). (See Wolak 1989a, 1991 for tests of nonlinear inequalities.) The testing problem is also related to that of Gouriéroux, Holly, and Monfort (1982), King and Smith (1986), and Andrews (1998), where the alternative is constrained by inequalities. (See Goldberger 1992 for a nice discussion of the relation between the two testing problems.)

An immediate consequence of Corollary 1 is that the RC is easy to manipulate by including *irrelevant alternative models*. This follows because the RC's p value, which is based on $\max(Z_1, \ldots, Z_m)$, can be increased in an artificial way by adding poor forecasts to the set of alternative forecasts (i.e., by increasing m while m_0 remains constant). In other words, it is possible to erode the power of the RC to 0 by including poor

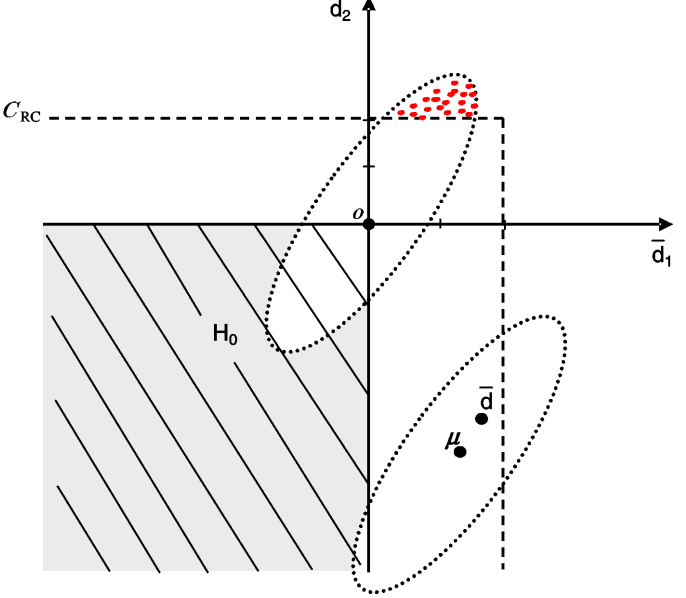

Figure 2. A Situation Where the RC Fails to Reject a False Null Hypothesis. The true parameter value is $\boldsymbol{\mu} = (\mu_1, \mu_2)'$, the sample estimate is $\bar{\mathbf{d}} = (\bar{d}_1, \bar{d}_2)'$, and C_{RC} represents the critical value derived from a null distribution that tacitly assumes that $\boldsymbol{\mu} = (0, 0)'$.

alternatives in the analysis. Naturally, we would want to avoid such properties to the extent possible.

Because the test statistics have asymptotic distributions that depend on $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$, these are nuisance parameters. The traditional way to proceed in this case is to substitute a consistent estimator for $\boldsymbol{\Omega}$ and use the LFC over the values of $\boldsymbol{\mu}$ that satisfy the null hypothesis. In the present situation, the point least favorable to the alternative is $\boldsymbol{\mu} = \mathbf{0}$, which presumes that all alternatives are as good as the benchmark. In the next section we explore an alternative way to handle the nuisance dependence on $\boldsymbol{\mu}$, where we use a data-dependent choice for $\boldsymbol{\mu}$ rather than $\boldsymbol{\mu} = \mathbf{0}$ as dictated by the LFC.

Figure 2 illustrates a situation for $m = 2$, where the two-dimensional plane represents the sampling space for $\bar{\mathbf{d}} = (\bar{d}_1, \bar{d}_2)'$. We have plotted a realization of $\bar{\mathbf{d}}$, that is in the neighborhood of its true expected value, $\boldsymbol{\mu} = (\mu_1, \mu_2)'$, and the ellipse around $\boldsymbol{\mu}$ is meant to illustrate the covariance structure of $\bar{\mathbf{d}}$. The shaded area represents the values of $\boldsymbol{\mu}$ that conform with the null hypothesis. Because we have placed $\boldsymbol{\mu}$ outside this shaded area, the situation in Figure 2 is one where the null hypothesis is false. The RC is an LFC-based test, so it derives critical values as if $\boldsymbol{\mu} = \mathbf{0}$ [the origin, $\mathbf{o} = (0, 0)'$, of the figure]. The critical value, C_{RC}, is represented by the dashed line, such that the area above and to the right of the dashed line defines the critical region of the RC. The shape of the critical region follows from the definition of T_n^{RC}. Because $\bar{\mathbf{d}}$ is outside the critical region in this example, the RC fails to reject the false null hypothesis in this case.

2.4 The Distribution Under the Null Hypothesis

Hansen (2003) proposed an alternative to the LFC approach that leads to more powerful tests of composite hypotheses. The LFC is based on a supremum taken over the null hypothesis, whereas the idea of Hansen (2003) is to take the supremum over a smaller (confidence) set chosen such that it contains the true parameter with a probability that converges to 1. In this article, we use a closely related procedure based directly on the asymptotic distributions of Theorem 1 and Corollary 1.

In the preceding section, we saw that the poor alternatives are irrelevant for the asymptotic distribution. So a proper test should reduce the influence of these models while preserving the influence of the models with $\mu_k = 0$. It may be tempting to simply exclude the alternatives with $\bar{d}_k < 0$ from the analysis. But this approach does not lead to valid inference in general, because the models that are (or appear to be) a little worse than the benchmark can have a substantial influence on the distribution of the test statistic in finite samples (and even asymptotically if $\mu_k = 0$). So we construct our test in a way that incorporates all models, while reducing the influence of alternatives that the data suggest are poor.

Our choice of estimator, $\hat{\boldsymbol{\mu}}^c$, is motivated by the law of the iterated logarithm stating that

$$P\left(\liminf_{n\to\infty} \frac{n^{1/2}(\bar{d}_k - \mu_k)}{\omega_k} = -\sqrt{2\log\log n}\right) = 1$$

and

$$P\left(\limsup_{n\to\infty} \frac{n^{1/2}(\bar{d}_k - \mu_k)}{\omega_k} = +\sqrt{2\log\log n}\right) = 1.$$

The first equality shows that $\hat{\mu}_k^c$ effectively captures all of the elements of $\boldsymbol{\mu}$ that are 0, such that $\mu_k = 0 \Rightarrow \hat{\mu}_k^c = 0$ almost surely. Similarly, if $\mu_k < 0$, then the second equality states that $\bar{d}_k$ is very close to μ_k; in fact, $n^{1/2}\bar{d}_k$ is smaller than $-n^{1/2-\epsilon}$ for any $\epsilon > 0$ and n sufficiently large. Thus $n^{1/2}\bar{d}_k/\omega_k$ is, in particular, smaller than the threshold rate, $-\sqrt{2\log\log n}$, for n sufficiently large, demonstrating that $\bar{d}_k$ eventually will stay below the implicit threshold in our definition of $\hat{\mu}_k^c$, such that $\mu_k < 0 \Rightarrow \hat{\mu}_k^c \ll 0$ almost surely. So $\hat{\boldsymbol{\mu}}^c$ meets the necessary asymptotic requirements that we identified in Theorem 1 and Corollary 1.

Although the poor alternatives should be discarded asymptotically, this is not the case in finite samples, as we discussed earlier. Our estimator, $\hat{\boldsymbol{\mu}}^c$, explicitly accounts for this by keeping all alternatives in the analysis. A poor alternative, $\mu_k < 0$, has an impact on the critical value whenever $\mu_k/(\omega_k n^{1/2})$ is only moderately negative, say between -1 and 0. This is the reason that *the poorly performing alternatives cannot simply be omitted from the analysis*. We emphasize this point because an earlier version of this article has been incorrectly quoted for "discarding the poor models."

Although $\hat{\boldsymbol{\mu}}^c$ leads to a correct separation of good and poor alternatives, other threshold rates also produce valid tests. The rate $\sqrt{2\log\log n}$ is the slowest rate that captures all alternatives with $\mu_k = 0$, whereas the faster rate, $n^{1/2-\epsilon}$ for any $\epsilon > 0$, guarantees that all of the poor models are discarded asymptotically. So a range of rates can be used to asymptotically discriminate between good and poor alternatives. One example is $\frac{1}{4}n^{1/4}$, which was used in a previous version of this article. Because different threshold rates will lead to different p values in finite samples, it is convenient to determine an upper and lower bound for the p values in which different threshold rates can result.

These are easily obtained using the "estimators," $\hat{\boldsymbol{\mu}}^l$ and $\hat{\boldsymbol{\mu}}^u$, given by $\hat{\mu}_k^l \equiv \min(\bar{d}_k, 0)$ and $\hat{\mu}_k^u = 0$, $k = 1, \ldots, m$, where the latter yields the LFC-based test. It is simple to verify that $\hat{\boldsymbol{\mu}}^l \leq \hat{\boldsymbol{\mu}}^c \leq \hat{\boldsymbol{\mu}}^u$, which in part motivates the superscripts, and we have the following result, where F_0 is the cdf of $\varphi(\mathbf{Z}, \mathbf{v}_0)$ that we defined in Theorem 1.

Theorem 2. Let F_n^i be the cdf of $\varphi(n^{1/2}\mathbf{Z}_n^i, \mathbf{V}_n)$, for $i = l, c$, or u, where $n^{1/2}(\mathbf{Z}_n^i - \hat{\boldsymbol{\mu}}^i) \xrightarrow{d} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Omega})$. Suppose that Assumptions 1 and 2 hold; then $F_n^c \to F_0$ as $n \to \infty$, for all continuity points of F_0 and $F_n^l(x) \leq F_n^c(x) \leq F_n^u(x)$ for all n and all $x \in \mathbb{R}$.

Theorem 2 demonstrates that $\hat{\boldsymbol{\mu}}^c$ leads to a consistent estimate of the asymptotic distribution of our test statistic. The theorem also demonstrates that $\hat{\boldsymbol{\mu}}^l$ and $\hat{\boldsymbol{\mu}}^u$ provide upper and lower bound for the distribution F_n^c that can be useful in practice; for example, a substantial difference between these bounds is indicative of the presence of poor alternatives, in which case the sample-dependent null distribution is useful.

Given a value for the test statistic $t = T_n(\mathbf{d}_1, \ldots, \mathbf{d}_n)$, it is natural to define the *true asymptotic p value* as $p_0(t) \equiv 1 - F_0(t)$. The empirical p value is deduced from an estimate of F_n^i, $i = l, c, u$, and the following corollary demonstrates that $\hat{\boldsymbol{\mu}}^c$ yields a consistent p value.

Corollary 2. Consider the studentized test statistic, $t = T_n^{\mathrm{SPA}}(\mathbf{d}_1, \ldots, \mathbf{d}_n)$. Let the empirical p value, $\hat{p}_n^c(t)$, be inferred from $\hat{F}_n^c$, where $\hat{F}_n^c(t) - F_n^c(t) = o(1)$ for all t. Then $\hat{p}_n^c(t) \xrightarrow{p} p_0(t)$ for any $t > 0$.

The two other choices, $\hat{\boldsymbol{\mu}}^l$ and $\hat{\boldsymbol{\mu}}^u$, do not produce consistent p values in general. It follows directly from Theorem 1 that $\hat{\boldsymbol{\mu}}^u$ will not produce a consistent p value unless $\boldsymbol{\mu} = \mathbf{0}$. That the p value from using $\hat{\boldsymbol{\mu}}^l$ is inconsistent is easily understood by noting that a critical value based on $\mathrm{N}_m(\mathbf{0}, \boldsymbol{\Omega})$ will be greater than one based on the mixed Gaussian distribution, $\mathrm{N}_m(n^{1/2}\hat{\boldsymbol{\mu}}^l, \boldsymbol{\Omega})$. So a p value based on $\hat{\boldsymbol{\mu}}^l$ is (asymptotically) smaller than the correct p value, which makes this a liberal test despite the fact that $\hat{\boldsymbol{\mu}}^l \xrightarrow{p} \boldsymbol{\mu}$ under the null hypothesis. This problem is closely related to the inconsistency of the bootstrap, when a parameter is on the boundary of the parameter space, as analyzed by Andrews (2000). In our situation the inconsistency arises because $\boldsymbol{\mu}$ is on the boundary of the null hypothesis, which leads to a violation of a *similarity on the boundary* condition (see Hansen 2003). (See Cox and Hinkley 1974, p. 150, and Gouriéroux and Monfort 1995, chap. 16, for discussions of the finite-sample version of this similarity condition.)

Figure 3 shows how the consistent estimate of the null distribution can improve the power. Recall the situation from Figure 2, where the null hypothesis is false. The data-dependent null distribution is defined from a projection of $\bar{\mathbf{d}} = (\bar{d}_1, \bar{d}_2)'$ onto the set of parameter values that conform with the null hypothesis. This yields the point a, which represents $\hat{\boldsymbol{\mu}}^l = \hat{\boldsymbol{\mu}}^c$ (assuming that $\bar{d}_2$ is below the relevant $2 \log \log n$-threshold). The critical region of the SPA test (induced by $\bar{\mathbf{d}}$) is the area above and to the right of the dotted line marked by C_{SPA}. Because $\bar{\mathbf{d}}$ is in the critical region, the SPA test (correctly) rejects the null hypothesis in this case.

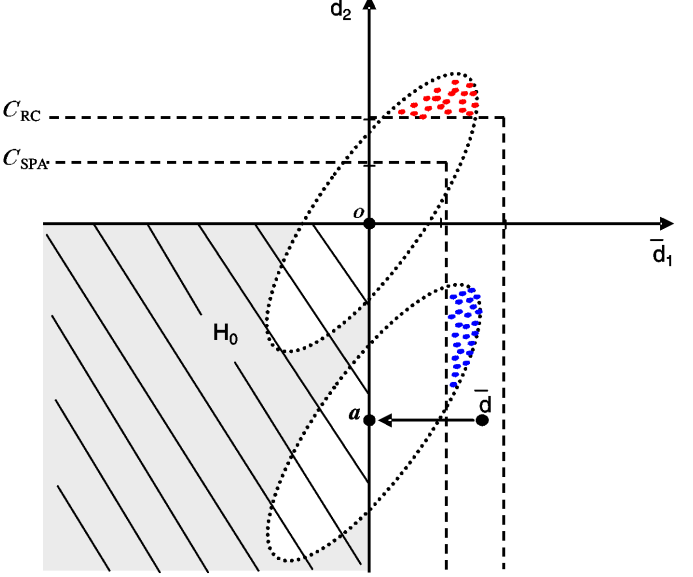

Figure 3. How the Power Is Improved by Using the Sample-Dependent Null Distribution. This distribution is centered about $\hat{\boldsymbol{\mu}}^c = a$, which leads to the critical value C_{SPA}. In contrast, the RC fails to reject the null hypothesis, because the LFC-based null distribution leads to the larger critical value C_{RC}.

3. BOOTSTRAP IMPLEMENTATION OF THE TEST FOR SUPERIOR PREDICTIVE ABILITY

In this section we describe a bootstrap implementation of the SPA tests in detail. The implementation is based on the stationary bootstrap of Politis and Romano (1994), but it is straightforward to modify the implementation to the block bootstrap of Künsch (1989). Although there are arguments that favor the block bootstrap over the stationary bootstrap (see Lahiri 1999), these advantages require the use of an optimal block length that is difficult to determine when m is large relative to n, as will often be the case when testing for SPA.

The stationary bootstrap of Politis and Romano (1994) is based on pseudo-time series of the original data. The pseudo-time series $\{\mathbf{d}_{b,t}^*\} \equiv \{\mathbf{d}_{\tau_{b,t}}\}$, $b = 1, \ldots, B$, are resamples of $\mathbf{d}_t$, where $\{\tau_{b,1}, \ldots, \tau_{b,n}\}$ is constructed by combining blocks of $\{1, \ldots, n\}$ with random lengths. The leading case is that where the block length is chosen to be geometrically distributed with parameter $q \in (0, 1]$, but the block length may be randomized differently, as discussed by Politis and Romano (1994). The number of bootstrap resamples, B, should be chosen to be sufficiently large such that the results are not affected by the actual draws of $\tau_{b,t}$. This can be achieved by increasing B until the results are robust to increments, or more formal methods, such as the three-step method of Andrews and Buchinsky (2000), can be applied. Here we follow the conventional setup of the stationary bootstrap and generate B resamples from two random $B \times n$ matrices, $\mathbf{U}$ and $\mathbf{V}$, where the elements, $u_{b,t}$ and $v_{b,t}$, are independent and uniformly distributed on $(0, 1]$. The first element of each resample is defined by $\tau_{b,1} = \lceil n u_{b,1} \rceil$, where $\lceil x \rceil$ is the smallest integer that is larger than or equal to x. For $t = 2, \ldots, n$, the elements are given recursively by

$$\tau_{b,t} = \begin{cases} \lceil n u_{b,1} \rceil & \text{if } v_{b,t} < q \\ \mathbb{1}_{\{\tau_{b,t-1} < n\}} \tau_{b,t-1} + 1 & \text{if } v_{b,t} \geq q. \end{cases}$$

So with probability q, the tth element is chosen uniformly on $\{1, \ldots, n\}$ and with probability $1-q$, the tth element is chosen to be the integer that follows $\tau_{b,t-1}$, unless $\tau_{b,t-1} = n$ in which case $\tau_{b,t} \equiv 1$. The block bootstrap is very similar to the stationary bootstrap, but instead of using blocks with random length, the block bootstrap combines blocks of equal length.

From the pseudo-time series, we calculate their sample averages, $\bar{\mathbf{d}}_b^* \equiv n^{-1} \sum_{t=1}^n \mathbf{d}_{b,t}^*$, $b = 1, \ldots, B$, that can be viewed as (asymptotically) independent draws from the distribution of $\bar{\mathbf{d}}$, under the bootstrap distribution. So this provides an intermediate step to estimate the distribution of our test statistic.

Lemma 1. Let Assumption 1 hold and suppose that the bootstrap parameter, $q = q_n$, satisfies $q_n \to 0$ and $nq_n^2 \to \infty$ as $n \to \infty$. Then

$$\sup_{\mathbf{z} \in \mathbb{R}^m} \left| P^*\left(n^{1/2}(\bar{\mathbf{d}}_b^* - \bar{\mathbf{d}}) \le \mathbf{z}\right) - P\left(n^{1/2}(\bar{\mathbf{d}} - \boldsymbol{\mu}) \le \mathbf{z}\right) \right| \xrightarrow{p} 0,$$

where P^* denotes the bootstrap probability measure.

This lemma demonstrates that the empirical distribution of the pseudo-time series can be used to approximate the distribution of $n^{1/2}(\bar{\mathbf{d}} - \boldsymbol{\mu})$. This result follows directly from Goncalves and de Jong (2003, thm. 2) who derived the result under slightly weaker assumptions than we have stated. (Their assumptions are formulated for near–epoch-dependent processes.) The test statistic T_n^{SPA} requires estimates of ω_k^2, $k = 1, \ldots, m$. An earlier version of this article was based on the estimator

$$\hat{\omega}_{k,B}^{*2} \equiv B^{-1} \sum_{b=1}^{B} \left(n^{1/2}\bar{d}_{k,b}^* - n^{1/2}\bar{d}_k\right)^2,$$

where $\bar{d}_{k,b}^* = n^{-1} \sum_{t=1}^n d_{k,\tau_{b,t}}$. By the law of large numbers, this estimator is consistent for the bootstrap population value of the variance, which in turn is consistent for the true variance, ω_k^2 (see Goncalves and de Jong 2003, thm. 1). However, it is our experience that B needs to be quite large to sufficiently reduce the additional layer of randomness introduced by the resampling scheme. So our recommendation is to use the bootstrap population value directly, which is given by

$$\hat{\omega}_k^2 \equiv \hat{\gamma}_{0,k} + 2 \sum_{i=1}^{n-1} \kappa(n,i) \hat{\gamma}_{i,k},$$

where

$$\hat{\gamma}_{i,k} \equiv n^{-1} \sum_{j=1}^{n-i} (d_{k,j} - \bar{d}_k)(d_{k,j+i} - \bar{d}_k), \qquad i = 0, 1, \ldots, n-1,$$

are the usual empirical covariances and the kernel weights (under the stationary bootstrap) are given by

$$\kappa(n,i) \equiv \frac{n-i}{n}(1-q)^i + \frac{i}{n}(1-q)^{n-i}$$

(see Politis and Romano 1994).

We seek the distribution of the test statistics under the null hypothesis, so we impose the null by recentering the bootstrap variables about $\hat{\boldsymbol{\mu}}^l$, $\hat{\boldsymbol{\mu}}^c$, or $\hat{\boldsymbol{\mu}}^u$. This is done by defining

$$Z_{k,b,t}^* \equiv d_{k,b,t}^* - g_i(\bar{d}_k),$$

$$i = l, c, u, b = 1, \ldots, B, t = 1, \ldots, n,$$

where $g_l(x) = \max(0, x)$, $g_c(x) = x \cdot \mathbb{1}_{\{x \ge -\sqrt{(\hat{\omega}_k^2/n) 2 \log\log n}\}}$, and $g_u(x) = x$. It is simple to verify that the expected values of $Z_{k,b,t}^*$, $i = l, c, u$ (conditional on $\mathbf{d}_1, \ldots, \mathbf{d}_n$), are given by $\hat{\boldsymbol{\mu}}^l$, $\hat{\boldsymbol{\mu}}^c$, and $\hat{\boldsymbol{\mu}}^u$.

Corollary 3. Let Assumption 1 hold and let $\mathbf{Z}_{b,t}^*$ be centered about $\hat{\boldsymbol{\mu}}$, for $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^l$, $\hat{\boldsymbol{\mu}}^c$, or $\hat{\boldsymbol{\mu}}^u$. Then

$$\sup_{\mathbf{z} \in \mathbb{R}^m} \left| P^*\left(n^{1/2}(\overline{\mathbf{Z}}_b^* - \hat{\boldsymbol{\mu}}) \le \mathbf{z}\right) - P\left(n^{1/2}(\bar{\mathbf{d}} - \boldsymbol{\mu}) \le \mathbf{z}\right) \right| \xrightarrow{p} 0,$$

where $\bar{Z}_{k,b}^* = n^{-1} \sum_{t=1}^n Z_{k,b,t}^*$, $k = 1, \ldots, m$.

Given our assumptions about the test statistic, Corollary 3 demonstrates that we can approximate the distribution of our test statistics under the null hypothesis by the empirical distribution we obtain from the bootstrap resamples $\mathbf{Z}_{b,t}^*$, $t = 1, \ldots, n$. The p values of the three tests for SPA are now simple to obtain. We calculate $T_{b,n}^{\mathrm{SPA}*} = \max\{0, \max_{k=1,\ldots,m}[n^{1/2}\bar{Z}_{k,b}^*/\hat{\omega}_k]\}$ for $b = 1, \ldots, B$, and the bootstrap p value is given by

$$\hat{p}_{\mathrm{SPA}} \equiv \sum_{b=1}^{B} \frac{\mathbb{1}_{\{T_{b,n}^{\mathrm{SPA}*} > T_n^{\mathrm{SPA}}\}}}{B},$$

where the null hypothesis should be rejected for small p values. Thus we obtain three p values, one for each of the estimators $\hat{\boldsymbol{\mu}}^l$, $\hat{\boldsymbol{\mu}}^c$, and $\hat{\boldsymbol{\mu}}^u$. The p values based on the test statistic T_n^{RC} can be derived similarly.

Note that we are using the same estimate of ω_k^2 to calculate T_n^{SPA} and $T_{b,n}^{\mathrm{SPA}*}$, $b = 1, \ldots, B$. A nice robustness property of the SPA test is that it is valid even if $\hat{\omega}_k^2$ is inconsistent for ω_k^2. This is easy to understand by recalling that $\hat{\omega}_k^2 = 1$ for all k leads to the RC (and 1 is generally inconsistent for ω_k^2). Although this robustness is convenient, it is desirable that $(\hat{\omega}_1^2, \ldots, \hat{\omega}_m^2)$ be close to $(\omega_1^2, \ldots, \omega_m^2)$, such that the individual statistics, $n^{1/2}\bar{d}_k/\hat{\omega}_k$, have approximately the same scale, due to the power issues that we discussed in Section 2.

4. SIZE AND POWER COMPARISON BY MONTE CARLO SIMULATIONS

The two test statistics T_n^{RC} and T_n^{SPA} and the three null distributions centered about $\hat{\boldsymbol{\mu}}^l$, $\hat{\boldsymbol{\mu}}^c$, and $\hat{\boldsymbol{\mu}}^u$ result in six different tests. In this section we study the size and power properties of these tests in a Monte Carlo experiment.

We generate $L_{k,t} \sim \text{iid}\, \mathrm{N}(\lambda_k/\sqrt{n}, \sigma_k^2)$ for $k = 0, 1, \ldots, m$ and $t = 1, \ldots, n$, where the benchmark model has $\lambda_0 = 0$. So positive values ($\lambda_k > 0$) correspond to alternatives that are worse than the benchmark, whereas negative values ($\lambda_k < 0$) correspond to alternatives that are better than the benchmark.

In our experiment we have $\lambda_1 \le 0$ and $\lambda_k \ge 0$ for $k = 2, \ldots, m$, such that the first alternative ($k = 1$) defines whether the rejection probability corresponds to a type I error ($\lambda_1 = 0$) or a power ($\lambda_1 < 0$). The performances of the "poor" models are such that their mean values are spread evenly between 0 and $\lambda_m = \Lambda_0$ (the worst model). So the vectors of the λ_k's are

given by

$$\boldsymbol{\lambda} \equiv \begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \vdots \\ \lambda_{m-1} \\ \lambda_m \end{pmatrix} = \begin{pmatrix} 0 \\ \Lambda_1 \\ \frac{1}{m-1}\Lambda_0 \\ \frac{2}{m-1}\Lambda_0 \\ \vdots \\ \frac{m-2}{m-1}\Lambda_0 \\ \Lambda_0 \end{pmatrix}.$$

In our experiments we use $\Lambda_0 = 0, 1, 2, 5, 10$ to control the extent to which the inequalities are binding with ($\Lambda_0 = 0$ corresponding to the case where all inequalities are binding). The first alternative model has $\Lambda_1 = 0, -1, -2, -3, -4, -5$. So $\lambda_1 = \Lambda_1$ defines the local alternative that is being analyzed (unless $\Lambda_1 = 0$, which conforms with the null hypothesis). To make the experiment more realistic, we tie the variance, σ_k^2, to the "quality" of the model. Specifically, we set

$$\sigma_k^2 = \tfrac{1}{2}\exp(\arctan(\lambda_k)),$$

such that a good model has a smaller variance than poor model. Note that this implies that

$$\sqrt{n}\bar{d}_k \sim \mathrm{N}(\mu_k, \omega_k^2),$$

where

$$\mu_k = \frac{\lambda_k}{\sqrt{n}} \qquad \text{and} \qquad \omega_k^2 \simeq 1 + \frac{1}{2}\lambda_k + \frac{1}{4}\lambda_k^2 - \frac{1}{12}\lambda_k^3,$$

where the expression for ω_k^2 now follows from $\mathrm{var}(d_{k,t}) = \mathrm{var}(L_{0,t} - L_{k,t}) = \frac{1}{2} + \mathrm{var}(L_{k,t})$ and the Taylor expansion (about 0)

$$\frac{1}{2}\exp(\arctan(x)) = \frac{1}{2}\left[1 + x + \frac{1}{2}x^2 - \frac{1}{6}x^3 - \frac{7}{24}x^4 + O(x^5)\right].$$

4.1 Simulation Results

First, we consider the case with $m = 100$ and the two sample sizes $n = 200$ and $n = 1{,}000$. Then we consider the case with $m = 1{,}000$ using the sample size $n = 200$. The rejection frequencies that we report are based on 10,000 independent samples, where we used $q = 1$ in accordance with the lack of time dependence in $\mathbf{d}_t$, $t = 1, \ldots, n$. All of our simulations were made using Ox 3.30 (see Doornik 1999). The rejection frequencies of the tests at the 5% and 10% levels are reported in Tables 2–4. The numbers in italic type are used when the null hypothesis is true ($\Lambda_1 = 0$), so these frequencies correspond to type I errors. The numbers in regular type represent powers for the various local alternatives ($\Lambda_1 < 0$).

Table 2. Rejection Frequencies Under the Null and Alternative ($m = 100$ and $n = 200$)

	Level: $\alpha = .05$						*Level: $\alpha = .10$*					
Λ_1	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u
Panel A: $\Lambda_0 = 0$												
0	*.055*	*.053*	*.053*	*.062*	*.060*	*.060*	*.108*	*.101*	*.101*	*.116*	*.110*	*.109*
−1	.057	.054	.054	.077	.074	.074	.112	.105	.105	.136	.129	.129
−2	.121	.111	.111	.310	.280	.280	.219	.197	.197	.436	.389	.388
−3	.550	.471	.470	.848	.764	.761	.727	.620	.618	.921	.845	.841
−4	.968	.888	.882	.997	.979	.976	.993	.947	.941	1.000	.990	.987
−5	1.000	.996	.992	1.000	1.000	1.000	1.000	.999	.998	1.000	1.000	1.000
Panel B: $\Lambda_0 = 1$												
0	*.013*	*.010*	*.010*	*.026*	*.022*	*.022*	*.035*	*.025*	*.025*	*.055*	*.044*	*.044*
−1	.013	.010	.010	.047	.041	.040	.036	.027	.027	.087	.072	.071
−2	.036	.028	.028	.312	.252	.250	.084	.060	.060	.436	.345	.342
−3	.301	.201	.197	.862	.744	.733	.516	.334	.327	.928	.829	.814
−4	.896	.677	.658	.998	.977	.971	.971	.816	.793	1.000	.989	.984
−5	1.000	.968	.952	1.000	1.000	.999	1.000	.991	.980	1.000	1.000	1.000
Panel C: $\Lambda_0 = 2$												
0	*.004*	*.002*	*.002*	*.018*	*.012*	*.012*	*.013*	*.007*	*.006*	*.039*	*.026*	*.026*
−1	.004	.002	.002	.044	.032	.032	.014	.007	.006	.080	.058	.056
−2	.013	.007	.006	.336	.244	.238	.041	.020	.019	.464	.336	.324
−3	.195	.077	.073	.881	.745	.721	.401	.167	.152	.941	.827	.799
−4	.842	.460	.414	.999	.978	.968	.957	.659	.598	1.000	.989	.982
−5	.999	.911	.855	1.000	1.000	.999	1.000	.971	.934	1.000	1.000	1.000
Panel D: $\Lambda_0 = 5$												
0	*.002*	*.000*	*.000*	*.014*	*.007*	*.005*	*.008*	*.001*	*.000*	*.032*	*.013*	*.011*
−1	.002	.000	.000	.056	.031	.025	.009	.001	.000	.101	.054	.044
−2	.012	.001	.001	.433	.273	.227	.047	.005	.003	.573	.370	.306
−3	.262	.032	.017	.929	.787	.710	.533	.088	.045	.968	.860	.784
−4	.913	.336	.167	1.000	.986	.966	.983	.581	.312	1.000	.995	.979
−5	1.000	.894	.620	1.000	1.000	.999	1.000	.974	.786	1.000	1.000	1.000
Panel E: $\Lambda_0 = 10$												
0	*.003*	*.000*	*.000*	*.016*	*.007*	*.002*	*.011*	*.001*	*.000*	*.036*	*.015*	*.006*
−1	.004	.000	.000	.080	.043	.022	.014	.001	.000	.149	.073	.039
−2	.037	.002	.000	.532	.340	.221	.128	.011	.001	.675	.455	.298
−3	.487	.064	.006	.953	.843	.703	.768	.181	.021	.980	.907	.779
−4	.973	.526	.091	1.000	.992	.964	.997	.772	.196	1.000	.998	.979
−5	1.000	.963	.462	1.000	1.000	.999	1.000	.993	.662	1.000	1.000	1.000

NOTE: Estimated rejection frequencies for the six tests for SPA under the null hypothesis ($\Lambda_1 = 0$) and local alternatives ($\Lambda_1 < 0$). The rejection frequencies in italic type correspond to type I errors, and those in regular type correspond to local powers. The reality check of White (2000) is denoted by RC_u, and the test advocated in this article is denoted by SPA_c.

Table 3. Rejection Frequencies Under the Null and Alternative ($m = 100$ and $n = 1{,}000$)

	Level: $\alpha = .05$						Level: $\alpha = .10$					
Λ_1	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u
Panel A: $\Lambda_0 = 0$												
0	*.051*	*.048*	*.048*	*.051*	*.048*	*.048*	*.104*	*.098*	*.098*	*.107*	*.100*	*.100*
−1	.054	.051	.051	.068	.064	.064	.110	.103	.103	.131	.122	.122
−2	.125	.116	.116	.309	.282	.282	.223	.202	.202	.435	.391	.390
−3	.556	.480	.479	.843	.762	.760	.729	.624	.622	.918	.842	.840
−4	.970	.889	.886	.998	.980	.977	.995	.945	.941	1.000	.992	.990
−5	1.000	.996	.994	1.000	1.000	1.000	1.000	.999	.997	1.000	1.000	1.000
Panel B: $\Lambda_0 = 1$												
0	*.011*	*.009*	*.009*	*.020*	*.017*	*.017*	*.031*	*.024*	*.023*	*.050*	*.040*	*.039*
−1	.011	.009	.009	.043	.036	.035	.033	.025	.025	.086	.069	.069
−2	.034	.026	.026	.312	.252	.250	.084	.059	.059	.436	.346	.342
−3	.316	.205	.203	.859	.740	.732	.520	.338	.331	.927	.822	.814
−4	.900	.682	.666	.999	.978	.972	.973	.816	.797	1.000	.990	.985
−5	1.000	.968	.955	1.000	1.000	.999	1.000	.991	.982	1.000	1.000	1.000
Panel B: $\Lambda_0 = 2$												
0	*.003*	*.001*	*.001*	*.014*	*.009*	*.009*	*.012*	*.004*	*.004*	*.034*	*.022*	*.021*
−1	.003	.002	.002	.042	.029	.028	.013	.004	.004	.079	.055	.054
−2	.014	.006	.006	.338	.242	.236	.042	.018	.017	.465	.330	.322
−3	.202	.082	.077	.881	.737	.720	.411	.169	.159	.941	.820	.798
−4	.844	.461	.428	.999	.979	.969	.959	.652	.602	1.000	.991	.983
−5	1.000	.906	.861	1.000	1.000	.999	1.000	.969	.936	1.000	1.000	1.000
Panel B: $\Lambda_0 = 5$												
0	*.002*	*.000*	*.000*	*.012*	*.005*	*.004*	*.006*	*.000*	*.000*	*.029*	*.011*	*.008*
−1	.002	.000	.000	.057	.028	.024	.007	.001	.000	.103	.051	.042
−2	.014	.001	.000	.435	.267	.225	.047	.004	.002	.572	.364	.306
−3	.270	.029	.017	.930	.777	.708	.540	.084	.044	.968	.851	.784
−4	.917	.328	.175	.999	.987	.966	.987	.554	.320	1.000	.995	.981
−5	1.000	.877	.632	1.000	1.000	.999	1.000	.966	.791	1.000	1.000	1.000
Panel B: $\Lambda_0 = 10$												
0	*.003*	*.000*	*.000*	*.013*	*.005*	*.003*	*.010*	*.001*	*.000*	*.033*	*.012*	*.005*
−1	.003	.000	.000	.083	.042	.022	.013	.001	.000	.145	.070	.039
−2	.039	.002	.000	.534	.335	.220	.128	.010	.000	.672	.444	.299
−3	.498	.060	.006	.954	.835	.703	.762	.165	.020	.980	.900	.778
−4	.974	.496	.095	.999	.994	.965	.997	.737	.203	1.000	.998	.980
−5	1.000	.953	.480	1.000	1.000	.999	1.000	.993	.669	1.000	1.000	1.000

NOTE: Estimated rejection frequencies for the six tests for SPA under the null hypothesis ($\Lambda_1 = 0$) and local alternatives ($\Lambda_1 < 0$). The rejection frequencies in italic type correspond to type I errors, and those in regular type correspond to local powers. The reality check of White (2000) is denoted by RC_u, and the test advocated in this article is denoted by SPA_c.

Table 2 presents the results for the case where $m = 100$ and $n = 200$. In the situation where all 100 inequalities are binding ($\Lambda_0 = \Lambda_1 = 0$), we see that the rejection probabilities are close to the nominal levels for all the tests. The SPA_c test has an overrejection by 1%. This overrejection appears to be a small-sample problem, because it disappears when the sample size is increased to $n = 1{,}000$ (see Table 3). The fact that the liberal null distribution does not lead to a larger overrejection is interesting. This finding may be due to the positive correlation across alternatives, $\operatorname{cov}(d_{i,t}, d_{j,t}) = \operatorname{var}(L_{0,t}) > 0$, which creates a positive correlation between the test statistic and $\hat{\boldsymbol{\mu}}^l$. Thus the critical value will tend to be (too) small, when the test statistic is small and this correlation will reduce the overrejection of the tests based on $\hat{\boldsymbol{\mu}}^l$. This suggests that our test may be improved if there is a reliable way to incorporate information about the off-diagonal elements of $\boldsymbol{\Omega}$. We do not pursue this aspect in this article.

Panel A corresponds to the case where $\boldsymbol{\mu} = \mathbf{0}$, that is, the best possible situation for LFC-based tests. This is the only situation where the LFC-based tests apply the correct asymptotic distribution, so it is not surprising that the tests based on $\hat{\boldsymbol{\mu}}^u = \mathbf{0}$ do well. Fortunately, our new test, SPA_c, also performs well in this case. Turning to the configurations where $\Lambda_0 > 0$, we immediately see the advantages of using the sample-dependent null distribution. A somewhat extreme situation is observed in Table 2, panel E for $(\Lambda_0, \Lambda_1) = (10, -3)$, whereas the RC almost never rejects the null hypothesis, while the new SPA_c-test has a power close to 84%.

Table 4 is quite interesting, because this is a situation where $m = 1{,}000$ exceeds the sample size $n = 200$. So in this case it is impossible to estimate $\boldsymbol{\Omega}$ in a sensible manner without imposing a restrictive structure on its coefficients. Thus using standard first-order asymptotics is not a viable alternative to the bootstrap implementation in this situation. Because the bootstrap invokes an implicit estimate of $\boldsymbol{\Omega}$, one might worry about its properties in this situation, where an explicit estimate is unavailable. Nevertheless, the bootstrap does surprisingly well, and we notice only a slight overrejection when all inequalities are binding ($\Lambda_0 = \Lambda_1 = 0$). The power properties are quite good despite the fact that 1,000 alternatives are being compared with the benchmark.

The power curves for the tests that use $\hat{\boldsymbol{\mu}}^c$ and $\hat{\boldsymbol{\mu}}^u$ are shown in Figure 4 for the case where $m = 100$, $n = 200$, and $\Lambda_0 = 20$. These power curves are based on tests that aim at a 5% significance level, and we have plotted their rejection frequencies against a range of local alternatives. These rejection frequencies have not been adjusted for their underrejection at $\Lambda_1 = 0$. This is a fair comparison, because it would not be possible to

Table 4. Rejection Frequencies Under the Null and Alternative ($m = 1{,}000$ and $n = 200$)

	Level: $\alpha = .05$						Level: $\alpha = .10$					
Λ_1	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u
Panel A: $\Lambda_0 = 0$												
0	*.049*	*.047*	*.047*	*.064*	*.062*	*.062*	*.106*	*.100*	*.100*	*.125*	*.119*	*.119*
−1	.049	.047	.047	.066	.064	.064	.106	.101	.100	.128	.122	.122
−2	.061	.058	.058	.173	.164	.164	.128	.121	.121	.269	.252	.252
−3	.288	.262	.262	.658	.598	.596	.434	.388	.388	.770	.699	.697
−4	.815	.720	.719	.980	.937	.933	.917	.828	.824	.994	.967	.963
−5	.998	.971	.967	1.000	.999	.998	1.000	.991	.988	1.000	1.000	1.000
Panel B: $\Lambda_0 = 1$												
0	*.009*	*.007*	*.007*	*.025*	*.022*	*.022*	*.022*	*.017*	*.017*	*.054*	*.045*	*.045*
−1	.009	.007	.007	.029	.025	.025	.022	.017	.017	.059	.050	.050
−2	.010	.008	.008	.150	.127	.127	.026	.020	.020	.229	.192	.191
−3	.066	.049	.049	.652	.555	.548	.150	.103	.102	.759	.652	.643
−4	.502	.345	.339	.980	.924	.916	.701	.500	.488	.993	.956	.947
−5	.965	.813	.794	1.000	.998	.997	.994	.907	.886	1.000	1.000	.999
Panel C: $\Lambda_0 = 2$												
0	*.001*	*.000*	*.000*	*.015*	*.011*	*.011*	*.005*	*.002*	*.002*	*.035*	*.026*	*.025*
−1	.001	.000	.000	.020	.015	.015	.005	.002	.002	.043	.032	.032
−2	.002	.000	.000	.155	.115	.113	.006	.003	.003	.233	.172	.167
−3	.016	.007	.007	.669	.544	.525	.054	.022	.022	.779	.636	.616
−4	.291	.125	.117	.985	.923	.906	.516	.243	.224	.994	.954	.940
−5	.901	.576	.529	1.000	.999	.996	.980	.744	.683	1.000	1.000	.998
Panel D: $\Lambda_0 = 5$												
0	*.000*	*.000*	*.000*	*.011*	*.005*	*.004*	*.002*	*.000*	*.000*	*.029*	*.012*	*.009*
−1	.000	.000	.000	.019	.010	.008	.002	.000	.000	.044	.020	.016
−2	.000	.000	.000	.199	.122	.101	.002	.000	.000	.291	.180	.148
−3	.011	.000	.000	.748	.570	.505	.045	.004	.002	.843	.664	.589
−4	.303	.036	.017	.993	.939	.897	.575	.098	.050	.998	.967	.930
−5	.936	.387	.207	1.000	.999	.996	.992	.605	.356	1.000	1.000	.998
Panel E: $\Lambda_0 = 10$												
0	*.001*	*.000*	*.000*	*.012*	*.004*	*.003*	*.002*	*.000*	*.000*	*.029*	*.011*	*.004*
−1	.001	.000	.000	.025	.012	.007	.002	.000	.000	.054	.024	.011
−2	.001	.000	.000	.259	.156	.097	.004	.000	.000	.366	.226	.141
−3	.031	.001	.000	.815	.633	.495	.109	.006	.000	.891	.726	.579
−4	.508	.064	.005	.996	.958	.892	.765	.175	.018	.999	.981	.926
−5	.983	.531	.099	1.000	1.000	.995	.998	.753	.210	1.000	1.000	.998

NOTE: Estimated rejection frequencies for the six tests for SPA under the null hypothesis ($\Lambda_1 = 0$) and local alternatives ($\Lambda_1 < 0$). The rejection frequencies in italic type correspond to type I errors, and those in regular type correspond to local powers. The reality check of White (2000) is denoted by RC_u, and the test advocated in this article is denoted by SPA_c.

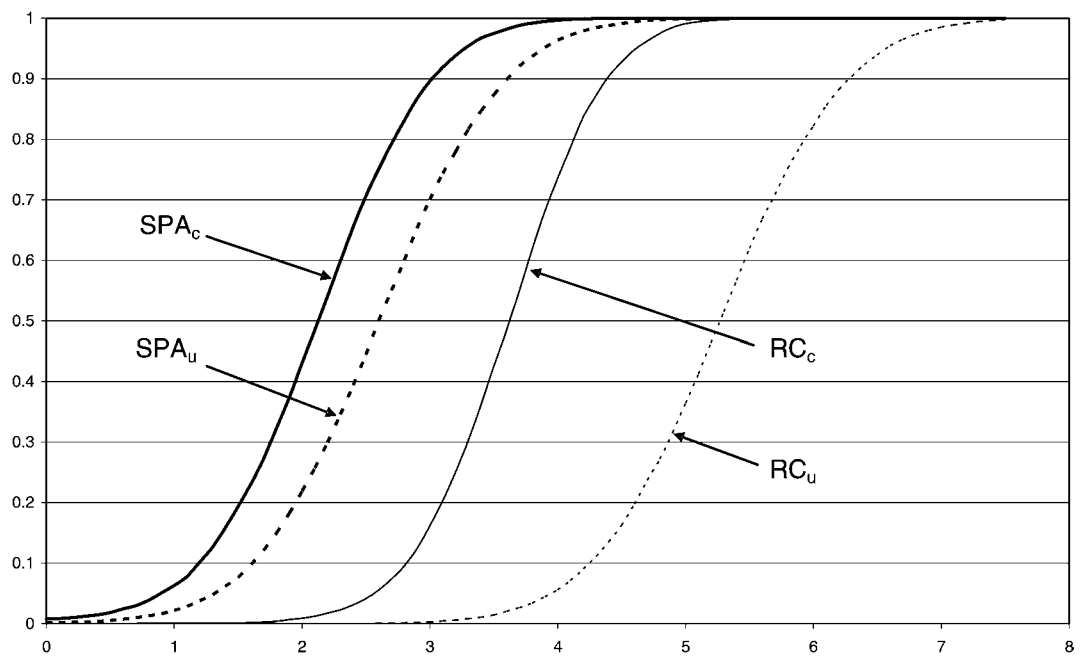

Figure 4. Local Power Curves of the Four Tests, SPA_c, SPA_u, RC_c, and RC_u, for the Simulation Experiment Where $m = 100$, $\Lambda_0 = 20$, and $\mu_1/\sqrt{n}$ $(= -\Lambda_1)$ Ranges From 0 to 8 (the x-axis). The power curves quantify the power improvements that are achieved by the two modifications of the reality check. Both the studentization and the data-dependent null distribution lead to substantial power gains in this design.

make such an adjustment in practice without exceeding the intended level of the test for other configurations—particularly the case where $\Lambda_0 = \Lambda_1 = 0$. (See Horowitz and Savin 2000 for a criticism of reporting "size"-adjusted powers.) From the power curves in Figure 4, it is clear that the RC is dominated by the three other tests. There is a substantial increase in power from using the consistent distribution, and a similar improvement is achieved by using the standardized test statistic, T_n^{SPA}. For example, the local alternative $\Lambda_1 = -4$ is rejected by the RC in about 5.5%. Using the data-dependent null distribution (RC_c) or the studentization (SPA_u) improves the power to about 73.6% and 96.4%. Invoking both modifications (as we advocate) improves the power to 99.7% in this simulation experiment. So both modifications are very useful, and the combination of the two yields a substantial improvement in power.

Comparing the sample sizes that would result in the same power is an effective way to convey the relative efficiency of the tests. For the configuration used in Figure 4, we see that the four tests have 50% power at the local alternatives, $\mu_1/\sqrt{n} \simeq 2.13$, 2.60, 3.63, and 5.28. These numbers demonstrate that we would need a sample size that is $(2.60/2.13)^2 = 1.49$ times larger to regain the power that is lost by using the LFC instead of the sample-dependent null distribution. In other words, using the LFC is equivalent to tossing away about 33% of the data. Similarly, dropping the studentization is equivalent to tossing away about 65% of the data, and dropping both modifications (i.e., using the RC instead of SPA_c) is equivalent to tossing away about 84% of the data in this simulation design.

5. FORECASTING U.S. INFLATION USING LINEAR REGRESSION MODELS

In an attempt to forecast annual U.S. inflation, we estimate a large number of linear regression models used to construct competing forecasts. The annual U.S. inflation rate is defined by $Y_t \equiv \log[P_t/P_{t-4}]$, where P_t is the GDP price deflator for the tth quarter. Inflation and most of the variables are not observed instantaneously. For this reason, we let the set of potential regressors consist of variables that are lagged five quarters or more relative to the end of the 12-month period for which we attempt to predict inflation. This leaves time (one quarter) for observing most of our regressors at the beginning of the 12-month period.

The linear regression models include 1, 2, or 3 regressors out of the pool of 27 regressors, $X_{1,t}, \ldots, X_{27,t}$, which leads to a total of 3,303 regression models. Descriptions and definitions of the regressors are given in Table 5.

The sequence of forecasts produced by the kth regression model is given by

$$\hat{Y}_{k,\tau+5} \equiv \hat{\boldsymbol{\beta}}'_{(k),\tau}\mathbf{X}_{(k),\tau}, \qquad \tau = 0, \ldots, n-1,$$

where $\mathbf{X}_{(k),\tau}$ contains the regressors included in model k and $\hat{\boldsymbol{\beta}}'_{(k),\tau}$ is the least squares estimator based on the 32 most recent observations (a rolling window). Thus $\hat{\boldsymbol{\beta}}_{(k),\tau} \equiv (\mathbf{X}'_{k,\tau}\mathbf{X}_{k,\tau})^{-1}\mathbf{X}'_{k,\tau}\mathbf{Y}_\tau$, where the rows of $\mathbf{X}_{k,\tau}$ are given by $\mathbf{X}'_{(k),t-5}$, $t = \tau - 32 + 1, \ldots, \tau$, and similarly the elements of $\mathbf{Y}_\tau$ are given by Y_τ, $t = \tau - 32 + 1, \ldots, \tau$. Using a rolling-window estimation scheme ensures that stationarity of $(\mathbf{X}_t, Y_t)$ is carried over to $L(Y_{t+h}, \hat{\boldsymbol{\beta}}'_{(k),t}\mathbf{X}_{(k),t})$, whereby a violation of Assumption 1 is avoided. For example, it is difficult to reconcile Assumption 1 with the case where $\boldsymbol{\beta}_{(k)}$ is estimated recursively (i.e., using an expanding window of observation as $n \to \infty$).

The first forecast of annual inflation is made at time 1959:Q4 (predicting inflation for the year 1960:Q1–1961:Q1). So the evaluation period includes $n = 160$ quarters:

$$t = \underbrace{1952{:}\mathrm{Q}1, \ldots, 1959{:}\mathrm{Q}4}_{\text{initial estimation period}}, \underbrace{1961{:}\mathrm{Q}1, \ldots, 2000{:}\mathrm{Q}4}_{\text{evaluation period}}.$$

The models are evaluated using a mean absolute error criterion (MAE) given by $L(Y_t, \hat{Y}_{k,t}) = |Y_t - \hat{Y}_{k,t}|$, and the best-performing models turn out to have a Phillips curve structure. In fact, the best forecasts are produced by regressors that measure (changes in) inflation, interest rates, employment, and GDP, and the very best sample performance is achieved by the three regressors $X_{1,t}$, $X_{8,t}$, and $X_{13,t}$, which represent annual inflation, employment relative to the previous year's employment, and the change in GDP (see Table 5). We also include the average forecast (i.e., average across all regression-based forecasts), because this simple combination of forecasts is often found to dominate the individual forecasts (see, e.g., Stock and Watson 1999). In addition to the average forecast, the 27 regressors lead to 3,303 regression-based forecasts when we consider all possible subset regressions with 1, 2, or 3 regressors. So we are to compare $m = 3{,}304$ forecasts to the random-walk benchmark, and we refer to this set of competing forecasts as the large universe.

Panel A of Table 6 contains the output produced by the tests for SPA for the large universe. Because the SPA_c p value is .741, there is no statistical evidence that any of the regression-based forecasts (including the average of them) are better than the random-walk forecast. Note the discrepancy between the p values based on $\hat{\boldsymbol{\mu}}^l$ and $\hat{\boldsymbol{\mu}}^u$. This difference suggests that some of the alternatives are poor forecasts, and a closer inspection of the large universe verifies that several models have substantially worse performance than the benchmark.

The ability to construct better forecasts using models with additional regressors is made difficult by the need to estimate additional parameters. In a forecasting exercise there is a trade-off between estimating a parameter and imposing it to have a particular value (typically 0, which is implicitly imposed on the coefficient of an omitted regressor). Imposing a particular value will (most likely) introduce a "bias," but if this bias is small, it may be less severe for out-of-sample predictions than the prediction error introduced by estimation error (see, e.g., Clements and Hendry 1998). Exploiting this bias–variance trade-off is particularly useful whenever the estimator is based on a moderate number of observations, as is the case in our application. For this reason, we also consider a small universe of regression-based forecasts, where all models include lagged inflation, $X_{1,t}$, as a predictor (regressor) with a coefficient set to unity. The remaining parameters are estimated by ridge regressions that shrink these parameters toward 0.

Thus the regression models have the form

$$Y_{\tau+5} - Y_\tau \equiv \boldsymbol{\beta}'_{(k)}\mathbf{X}_{(k),\tau} + \varepsilon_{(k),t}, \qquad \tau = 0, \ldots, n-1,$$

where $\mathbf{X}_{(k),\tau}$ is a vector that includes either one or two regressors. As before, we use a rolling-window scheme (32 quar-

Table 5. Definitions of Variables

Panel A: Description of variables

Y_t	Annual inflation
$X_{1,t}$, $X_{2,t}$	Annual inflation (lags of Y_t)
$X_{3,t}$, $X_{4,t}$	Quarterly inflation
$X_{5,t}$	Quarterly inflation relative to previous year's inflation
$X_{6,t}$, $X_{7,t}$	Changes in employment in manufacturing sector
$X_{8,t}$	Quarterly employment relative to average of previous year
$X_{9,t}$	Quarterly employment relative to average of previous 2 years
$X_{10,t}$, $X_{11,t}$	Quarterly changes in real inventory
$X_{12,t}$, $X_{13,t}$	Quarterly changes in quarterly GDP
$X_{14,t}$	Interest paid on 3-month T-bill
$X_{15,t}$, $X_{16,t}$	Changes in 3-month T-bill
$X_{17,t}$, $X_{18,t}$	Changes in 3-month T-bill relative to level of T-bill
$X_{19,t}$, $X_{20,t}$	Changes in prices of fuel and energy
$X_{21,t}$, $X_{22,t}$	Changes in prices of food
$X_{23,t}$–$X_{26,t}$	Quarterly dummies: first, second, third, and fourth quarters
$X_{27,t}$	Constant

Panel B: Definitions of variables

$Y_t = \log(\text{GDPCTPI}_t) - \log(\text{GDPCTPI}_{t-4}), \quad X_{1,t} = Y_{t-5}, \quad X_{2,t} = Y_{t-8}$

$X_{3,t} = 4[\log(\text{GDPCTPI}_t) - \log(\text{GDPCTPI}_{t-1}], \quad X_{4,t} = X_{3,t-1}$

$X_{5,t} = \log(1 + X_{3,t}) - \log(1 + X_{1,t-1})$

$X_{6,t} = \log(\text{MANEMP}_t) - \log(\text{MANEMP}_{t-1}), \quad X_{7,t} = X_{6,t-1}$

$X_{8,t} = \log(\text{MANEMP}_t) - \log(\frac{1}{4}\sum_{i=1}^{4} \text{MANEMP}_{t-\text{i}})$

$X_{9,t} = \log(\text{MANEMP}_t) - \log(\frac{1}{8}\sum_{i=1}^{8} \text{MANEMP}_{t-\text{i}})$

$X_{10,t} = \log(\text{CBI}_t) - \log(\text{GDP}_t), \quad X_{11,t} = X_{10,t-1}$

$X_{12,t} = \log(\text{GDP}_t) - \log(\text{GDP}_{t-1}), \quad X_{13,t} = X_{12,t-1}$

$X_{14,t} = \text{TB3MS}_t, \quad X_{15,t} = \Delta X_{14,t}, \quad X_{16,t} = X_{15,t-1}, \quad X_{17,t} = \Delta X_{14,t}/X_{14,t}, \quad X_{18,t} = X_{17,t-1}$

$X_{19,t} = \log(\text{PPIENG}_t) - \log(\text{PPIENG}_{t-1}), \quad X_{20,t} = X_{19,t-1}$

$X_{21,t} = \log(\text{PPIFCF}_t) - \log(\text{PPIFCF}_{t-1}), \quad X_{22,t} = X_{21,t-1}$

$X_{23,t} = 1, \quad X_{24,t} = X_{23,t-1}, \quad X_{25,t} = X_{23,t-2}, \quad X_{26,t} = X_{23,t-3}, \quad X_{27,t} = 1$

Raw data: GDPCTPI = gross domestic product: chain-type price index; CBI = change in private inventories; GDP = gross domestic product; TB3MS = 3-month Treasury bill rate, secondary market*; PPIENG = producer price index: fuels and related products and power**; PPIFCF = producer price index: finished consumer foods**; MANEMP = employees on nonfarm payrolls: manufacturing.

∗ Quarterly data are defined as the average of the monthly observations over the quarter.

∗∗ Quarterly data are defined as be the last monthly observation of the quarter.

Table 6. Tests for Superior Predictive Ability

				Loss	*t-statistic*	*p value*
Panel A: Results for the large universe of forecasts						
Evaluated by MAE			Benchmark:	.0098		
$m = 3{,}304$ (number of models)			Best performing:	.0084	1.2363	.120
$n = 160$ (sample size)			Most significant:	.0085	1.2628	.112
$B = 10{,}000$ (resamples)			Median:	.0141	−2.7694	
$q = .25$ (dependence)			Worst:	.0416	−7.8939	
	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u
SPA p values	.503	.781	.978	.571	**.741**	.903
Panel B: Results for the small universe of forecasts						
Evaluated by MAE			Benchmark:	.0098		
$m = 352$ (number of models)			Best performing:	.0082	2.7547	.006
$n = 160$ (sample size)			Most significant:	.0096	2.9399	.004
$B = 10{,}000$ (resamples)			Median:	.0097	.0657	
$q = .25$ (dependence)			Worst:	.0107	−1.3272	
	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u
SPA p values	.071	.106	.106	.045	**.048**	.048
Panel C: Results for the full universe of forecasts						
Evaluated by MAE			Benchmark:	.0098		
$m = 3{,}656$ (number of models)			Best performing:	.0082	2.7547	.006
$n = 160$ (sample size)			Most significant:	.0096	2.9399	.004
$B = 10{,}000$ (resamples)			Median:	.0135	−1.9398	
$q = .25$ (dependence)			Worst:	.0416	−7.8939	
	RC_l	RC_c	RC_u	SPA_l	SPA_c	SPA_u
SPA p value	.395	.691	.963	.078	**.100**	.135

NOTE: The table reports SPA p values for three sets of regression-based forecasts that are compared to a random-walk forecast. The p value of the new test, SPA_c, is in bold type. Panel A contains the results for the large universe, panel B contains the results for the small universe, and panel C contains the results for the full universe. For each "universe of forecasts" we also report the sample loss for the benchmark and the four alternative forecasts that had the smallest sample loss, the largest t-statistic for relative sample loss ($\bar{d}_k$), the median sample loss (across alternatives), and the worst sample loss. The corresponding t-statistic (of their sample loss relative to the benchmark) is given in the second last column. We also report the "p values" from the pairwise comparisons of "best" and "largest t-statistic" forecasts to the benchmark. These p values (unlike the SPA p values) do not account for the entire universe of forecasts.

ters), but with the estimator for $\boldsymbol{\beta}_{(k)}$ now given by $\hat{\boldsymbol{\beta}}_{(k),\tau} \equiv (\mathbf{X}'_{k,\tau}\mathbf{X}_{k,\tau} + \lambda\mathbf{I})^{-1}\mathbf{X}'_{k,\tau}\tilde{\mathbf{Y}}_\tau$, where $\lambda = .1$ is the shrinkage parameter and the elements of $\tilde{\mathbf{Y}}_\tau$ are given by $Y_t - Y_{t-5}$ for $t = \tau - 32 + 1, \ldots, \tau$. This results in 351 regression-based forecasts plus the average forecast, such that the total number of alternative forecasts in the small universe is $m = 352$. The most accurate forecast in the small universe is produced by the regression model with the regressors $X_{8,t}$ and $X_{9,t}$, which are two measures of (relative) employment. The largest t-statistic is produced by the regressions $X_{6,t}$ and $X_{10,t}$, which represent changes in employment and inventories. So our findings support a conclusion reached by Stock and Watson (1999) that forecasts based on Phillips curve specifications are useful for forecasting inflation.

The empirical results for this universe are presented in panel B of Table 6. The SPA_c p value for this universe is .048, which suggests that the benchmark is outperformed by the regression-based forecasts. For each of the test statistics, we note that the p values are quite similar. This agreement is not surprising, because the worst forecast is only slightly worse than the benchmark, such that $\hat{\boldsymbol{\mu}}^l$ and $\hat{\boldsymbol{\mu}}^u$ are similar. The difference in p values across the two test statistics is fairly modest but do suggest some variation in the variances, ω_k^2, $k = 1, \ldots, 352$.

Reporting the results in panel B is not fully consistent with the spirit of this article, because the results in panel B do not control for the 3,304 forecasting models that were compared to the benchmark in the initial analysis of the large universe. So the significant p values in panel B are subject to the criticisms of data mining. To address this concern, we perform the tests for SPA over the union of the large universe and the small universe. We refer to this set of forecasts as the full universe, and present the results for this set of alternatives in panel C. Adding the large number of insignificant alternatives to the comparison reduces the significance, although the excess performance continues to be borderline significant with an SPA_c p value of 10%. Note that the RC's p value increases from 10.6% to 96.3% by "adding" the large universe to the small universe. This jump in the p value is most likely due to the RC's sensitivity to poor and irrelevant alternatives. The SPA_c test's p value increases from 4.8% to 10%. Although this increment is more moderate, it reveals that the new test is not entirely immune to the inclusion of (a large number of) poor forecasts. This reminds us that excessive data mining can be costly in terms of the conclusions that can be drawn from an empirical analysis, because it may prevent the researcher from concluding that a particular finding is significant. Given the scarcity of macroeconomic data, it will often be useful to confine the set of alternatives to those motivated by theoretical considerations, instead of undertaking a blind search over a large number of alternatives.

6. SUMMARY AND CONCLUDING REMARKS

We have analyzed the problem of comparing multiple forecasts to a given benchmark through tests for superior predictive ability. We have shown that the power can be improved (often substantially) by using a studentized test statistic and incorporating additional sample information by means of a data-dependent null distribution. The latter serves to identify the irrelevant alternatives and reduce their influence on the test for SPA.

The power improvements were quantified in simulation experiments and an empirical forecasting exercise of U.S. inflation. These also highlighted that the RC is sensitive to poor and irrelevant alternatives. Two researchers are therefore more likely to arrive at the same conclusion when they use the SPA_c test than they would when using the RC—even if they do not fully agree on the set of forecasts that is relevant for the analysis.

Interestingly, we found that the best (and most significant) predictions of U.S. inflation were produced by regression-based forecasts that had a Phillips curve structure. In our full universe of alternatives, we found that the (random-walk) benchmark forecast is outperformed by the regression-based forecasts if a moderate (10%) significance level is used. Whereas the SPA_c test yields a p value of 10%, the RC yields a p value of about 96%, such that the two tests arrive at opposite conclusions (weak evidence against H_0 vs. no evidence). This is caused by the poor alternatives that conceal the evidence against the null hypothesis when the RC is used. This phenomenon was also discussed Hansen and Lunde (2005b), who compared a large number of volatility models using the methods of this article.

Although there are several advantages of our new test, some important issues need to be addressed in future research. In this article we have proposed two modifications and adopted these in a stationary framework. This framework does not permit the comparison of parameterized models when a recursive scheme is used to estimate the parameters. So it would be interesting to construct a suitable test that can accommodate this situation and analyze the need for our two modifications in this framework.

Despite its many pitfalls, data mining is a constructive device for the discovery of true phenomena and has become a popular tool in many applied areas, such as genetics, e-commerce, and financial services. However, it is necessary to account for the full data exploration before making a legitimate statement about significance. Increasing the number of comparisons raises the bar at which alternatives are classified as being significantly better than the benchmark. This aspect is particularly problematic for economic applications where data are scarce. In this context it is particularly useful to confine the exploration to alternatives motivated by theoretical considerations. Our empirical application provides a good illustration of this issue. Within the small universe we found fairly compelling evidence against the null hypothesis, and ex post it is easy to produce arguments that motivate the use of shrinkage methods, which led to the small universe of regression-based forecasts. However, because the large universe was explored in the initial analysis, we cannot exclude the possibility that the largest t-statistic would have been found in this universe. The weaker evidence against the null hypothesis found in the full universe is the price that we have to pay for the data exploration that preceded our analysis of the small universe.

ACKNOWLEDGMENTS

I thank Albert Chun, Jinyong Hahn, James D. Hamilton, Søren Johansen, Tony Lancaster, Asger Lunde, Michael

McCracken, Barbara Rossi, and seminar participants at Princeton, Harvard/MIT, University of Montreal, UBC, New York Fed, Stanford, NBER/NSF Summer Institute 2001, three anonymous referees, and Torben G. Andersen (editor) for many valuable comments and suggestions. I am also grateful for financial support from the Danish Research Agency (grant 24-00-0363).

APPENDIX: PROOFS

Proof of Theorem 1

We define the vectors, $\mathbf{W}_n, \mathbf{Z}_n \in \mathbb{R}^m$, whose elements are given by $W_{n,k} = n^{1/2}\bar{d}_k \mathbb{1}_{\{\mu_k<0\}}$ and $Z_{n,k} = n^{1/2}\bar{d}_k \mathbb{1}_{\{\mu_k=0\}}$, $k = 1, \ldots, m$. Thus $\mathbf{U}_n = \mathbf{W}_n + \mathbf{Z}_n$ under the null hypothesis. The mappings (coordinate selectors) that transform $\mathbf{U}_n$ into $\mathbf{W}_n$ and $\mathbf{Z}_n$ are continuous, so that $(\mathbf{W}_n - n^{1/2}\boldsymbol{\mu}) \xrightarrow{d} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Omega} - \boldsymbol{\Omega}^0)$ and $\mathbf{Z}_n \xrightarrow{d} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Omega}^0)$ by the continuous mapping theorem. This implies that

$$\begin{aligned}\varphi(\mathbf{U}_n, \mathbf{V}_n) &= \varphi(\mathbf{W}_n + \mathbf{Z}_n, \mathbf{V}_n) \\ &= \varphi(\mathbf{Z}_n, \mathbf{V}_n) + o_p(1) \xrightarrow{d} \varphi(\mathbf{Z}, \mathbf{v}_0),\end{aligned}$$

where the second equality uses Assumption 2(b) and the fact that the elements of $\mathbf{W}_n$ are either 0 ($\mu_k = 0$) or diverges to minus infinity in probability ($\mu_k < 0$). Under the alternative hypothesis, there will be an element of $n^{1/2}\bar{\mathbf{d}}$ that diverges to infinity. So the last result of the theorem follows by Assumption 2(c).

Proof of Corollary 1

The results follow from $n^{1/2}(\bar{\mathbf{d}} - \boldsymbol{\mu}) \xrightarrow{d} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Omega})$ and the continuous mapping theorem, applied to the mapping $\bar{\mathbf{d}} \mapsto \bar{d}_{\max}$.

Proof of Theorem 2

Without loss of generality, suppose that $\mu_k = 0$ for $k = 1, \ldots, m_0$ and that $\mu_k < 0$ for $k = m_0 + 1, \ldots, m$. Given our definition of $\hat{\boldsymbol{\mu}}^c$, it holds that $P(\hat{\mu}^c_1 = \cdots = \hat{\mu}^c_{m_0} = 0, \hat{\mu}^c_{m_0+1} < \epsilon, \ldots, \hat{\mu}^c_m < \epsilon)$ almost surely as $n \to 0$, for some $\epsilon < 0$. So for n sufficiently large, the last $m - m_0$ elements of $\mathbf{Z}^i_n$ are bounded below 0 in probability, which demonstrates that $\hat{\boldsymbol{\mu}}^c$ leads to the correct limiting distribution and $F^c_n \to F_0$. That $F^l_n(x) \le F^c_n(x) \le F^u_n(x)$ follows from $\hat{\boldsymbol{\mu}}^l \le \hat{\boldsymbol{\mu}}^c \le \hat{\boldsymbol{\mu}}^u$.

Proof of Corollary 2

The test statistic T^{SPA}_n leads to a continuous asymptotic distribution, $F_0(t)$, for all $t > 0$. Because $\hat{F}^c_n(t) - F_0(t) = [\hat{F}^c_n(t) - F^c_n(t)] + [F^c_n(t) - F_0(t)]$, the result now follows, because the first term is $o(1)$ by assumption and the second term is $o(1)$ by Theorem 2.

Proof of Lemma 1

This follows from work of Goncalves and de Jong (2003, thm. 2).

Proof of Corollary 3

Because $Z^*_{k,b,t} - \hat{\mu}_k = (d^*_{k,b,t} - g_i(\bar{d}_k)) - (\bar{d}_k - g_i(\bar{d}_k)) = d^*_{k,b,t} - \bar{d}_k$ for all $k = 1, \ldots, m$, the corollary follows trivially from Lemma 1.

[Received February 2005. Revised April 2005.]

REFERENCES

Andrews, D. W. K. (1998), "Hypothesis Testing With a Restricted Parameter Space," *Journal of Econometrics*, 84, 155–199.

——— (2000), "Inconsistency of the Bootstrap When a Parameter Is on the Boundary of the Parameter Space," *Econometrica*, 68, 399–405.

Andrews, D. W. K., and Buchinsky, M. (2000), "A Three-Step Method for Choosing the Number of Bootstrap Repetitions," *Econometrica*, 68, 23–52.

Chao, J. C., Corradi, V., and Swanson, N. R. (2001), "An Out-of-Sample Test for Granger Causality," *Macroeconomic Dynamics*, 5, 598–620.

Clark, T. E., and McCracken, M. W. (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics*, 105, 85–110.

Clements, M. P., and Hendry, D. F. (1998), *Forecasting Economic Time Series*, Cambridge, U.K.: Cambridge University Press.

Corradi, V., and Swanson, N. R. (2002), "A Consistent Test for Nonlinear Out-of-Sample Predictive Accuracy," *Journal of Econometrics*, 110, 353–381.

——— (2005a), "Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes," available at *http://econweb.rutgers.edu/nswanson/papers.htm*.

——— (2005b), "Predictive Density and Conditional Confidence Interval Accuracy Tests," *Journal of Econometrics*, forthcoming.

Corradi, V., Swanson, N. R., and Olivetti, C. (2001), "Predictive Ability With Cointegrated Variables," *Journal of Econometrics*, 104, 315–358.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman & Hall.

de Jong, R. (1997), "Central Limit Theorems for Dependent Heterogeneous Random Variables," *Econometric Theory*, 13, 353–367.

Diebold, F. X., and Mariano, R. S. (1995), "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263.

Doornik, J. A. (1999), *Ox: An Object-Oriented Matrix Language* (3rd ed.), London: Timberlake Consulants Press.

Dufour, J.-M. (1989), "Nonlinear Hypotheses, Inequality Restrictions, and Non-Nested Hypotheses: Exact Simultaneous Test in Linear Regressions," *Econometrica*, 57, 335–355.

Dufour, J.-M., and Khalaf, L. (2002), "Exact Tests for Contemporaneous Correlation of Disturbances in Seemingly Unrelated Regressions," *Journal of Econometrics*, 106, 143–170.

Folks, L. (1984), "Combination of Independent Tests," in *Handbook of Statistics 4: Nonparametric Methods*, eds. P. R. Krishnaiah and P. K. Sen, New York: North-Holland, pp. 113–121.

Giacomini, R., and White, H. (2003), "Tests of Conditional Predictive Ability," Working Paper 572, Boston College.

Goldberger, A. S. (1992), "One-Sided and Inequality Tests for a Pair of Means," in *Contributions to Consumer Demand and Econometrics*, eds. R. Bewley and T. V. Hoa, New York: St. Martin's Press, pp. 140–162.

Goncalves, S., and de Jong, R. (2003), "Consistency of the Stationary Bootstrap Under Weak Moment Conditions," *Economics Letters*, 81, 273–278.

Gouriéroux, C., Holly, A., and Monfort, A. (1982), "Likelihood Ratio Test, Wald Test, and Kuhn–Tucker Test in Linear Models With Inequality Constraints on the Regression Parameters," *Econometrica*, 50, 63–80.

Gouriéroux, C., and Monfort, A. (1995), *Statistics and Econometric Models*, Cambridge, U.K.: Cambridge University Press.

Hansen, P. R. (2003), "Asymptotic Tests of Composite Hypotheses," available at *http://www.stanford.edu/people/peter.hansen*.

Hansen, P. R., and Lunde, A. (2005a), "Consistent Ranking of Volatility Models," *Journal of Econometrics*, forthcoming.

——— (2005b), "A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1, 1)?" *Journal of Applied Econometrics*, forthcoming.

Harvey, D., and Newbold, P. (2000), "Tests for Multiple Forecast Encompassing," *Journal of Applied Econometrics*, 15, 471–482.

Harvey, D. I., Leybourne, S. J., and Newbold, P. (1997), "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting*, 13, 281–291.

——— (1998), "Tests for Forecast Encompassing," *Journal of Business & Economic Statistics*, 16, 254–259.

Horowitz, J. L., and Savin, N. E. (2000), "Empirically Relevant Critical Values for Hypothesis Tests: A Bootstrap Approach," *Journal of Econometrics*, 95, 375–389.

Inoue, A., and Kilian, L. (2004), "In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?" *Econometrics Review*, 23, 371–402.

King, M. L., and Smith, M. D. (1986), "Joint One-Sided Tests of Linear Regression Coefficients," *Journal of Econometrics*, 32, 367–387.

Künsch, H. R. (1989), "The Jackknife and the Bootstrap for General Stationary Observations," *The Annals of Statistics*, 17, 1217–1241.

Lahiri, S. N. (1999), "Theoretical Comparisons of Block Bootstrap Methods," *The Annals of Statistics*, 27, 386–404.

Lehmann, E. L., and Romano, J. P. (2005), *Testing Statistical Hypotheses* (3rd ed.), New York: Wiley.

Marden, J. I. (1985), "Combining Independent One-Sided Noncentral *t* or Normal Mean Tests," *The Annals of Statistics*, 13, 1535–1553.

McCracken, M. W. (2000), "Robust Out-of-Sample Inference," *Journal of Econometrics*, 99, 195–223.

Miller, R. G. (1981), *Simultaneous Statistical Inference* (2nd ed.), New York: Springer-Verlag.

Perlman, M. D. (1969), "One-Sided Testing Problems in Multivariate Analysis," *The Annals of Mathematical Statistics*, 40, 549–567.

Perlman, M. D., and Wu, L. (1999), "The Emperor's New Tests," *Statistical Science*, 14, 355–381.

Politis, D. N., and Romano, J. P. (1994), "The Stationary Bootstrap," *Journal of the American Statistical Association*, 89, 1303–1313.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), *Order-Restricted Statistical Inference*, New York: Wiley.

Romano, J. P., and Wolf, M. (2005), "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica*, forthcoming.

Savin, N. E. (1984), "Multiple Hypothesis Testing," in *Handbook of Econometrics*, Vol. 2, eds. K. J. Arrow and M. D. Intriligator, Amsterdam: North-Holland, pp. 827–879.

Stock, J. H., and Watson, M. W. (1999), "Forecasting Inflation," *Journal of Monetary Economics*, 44, 293–335.

Sullivan, R., Timmermann, A., and White, H. (1999), "Data-Snooping, Technical Trading Rules, and the Bootstrap," *Journal of Finance*, 54, 1647–1692.

——— (2001), "Dangers of Data-Driven Inference: The Case of Calendar Effects in Stock Returns," *Journal of Econometrics*, 105, 249–286.

——— (2003), "Forecast Evaluation With Shared Data Sets," *International Journal of Forecasting*, 19, 217–227.

Tippett, L. H. C. (1931), *The Methods of Statistics*, London: Williams and Norgate.

West, K. D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica*, 64, 1067–1084.

——— (2001), "Tests for Forecast Encompassing When Forecasts Depend on Estimated Regression Parameters," *Journal of Business & Economic Statistics*, 19, 29–33.

West, K. D., and McCracken, M. W. (1998), "Regression Based Tests of Predictive Ability," *International Economic Review*, 39, 817–840.

Westfall, P. H., and Young, S. S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustments*, New York: Wiley.

White, H. (2000), "A Reality Check for Data Snooping," *Econometrica*, 68, 1097–1126.

Wolak, F. A. (1987), "An Exact Test for Multiple Inequality and Equality Constrants in the Linear Regression Model," *Journal of the American Statistical Association*, 82, 782–793.

——— (1989a), "Local and Global Testing of Linear and Nonlinear Inequality Constraints in Nonlinear Econometric Models," *Econometric Theory*, 5, 1–35.

——— (1989b), "Testing Inequality Constraints in Linear Econometric Models," *Journal of Econometrics*, 41, 205–235.

——— (1991), "The Local Nature of Hypothesis Tests Involving Inequality Constraints in Nonlinear Models," *Econometrica*, 59, 981–995.